THE UNIVERSITY
of EDINBURGH

This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e.g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.

This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.

THE UNIVERSITY *of* EDINBURGH

A single cell transcriptomic analysis of autism associated gene expression in human cerebral cortex development

By Yifei Yang

Thesis submitted for the degree of Doctor of Philosophy

at the

School of Biomedical Sciences

University of Edinburgh

June 2020

Photo taken at The Meadows, 2019

Study at Edinburgh

January 2015 - November 2019

Disclaimer

I (Yifei Yang) performed all the bioinformatics analysis presented in this thesis unless otherwise clearly stated in the text. No part of this work has been or is being submitted for any other degree of qualification.

Signed:

Date: 26/06/2020

Acknowledgements

I would like to thank my supervisor, Dr. Thomas Pratt, and my co-supervisor, Prof. David J. Price, for their helpful discussions and other contributions which made my PhD thesis possible.

I would also like to thank my collaborators who generated and analysed the single-cell RNA sequencing data that are crucial for my PhD thesis, including Dr. Da Mi from Oscar Marin's lab, as well as Dr. Zhen Li from Nenad Sestan's lab.

Other colleagues in the lab are fundamental to my projects as well. Sarah Morson, a PhD student in my lab, has been collaborating with me in my project and help me explain the bioinformatics results I generated. Dr. James Clegg introduced me the knowledge of neuroscience when I started my PhD study. I would also thank Dr. Hannah Parkin and Dr. Calvin Chan for their kindly welcome when I join this lab.

I would also like to thank everyone that I have met in various seminars and conferences for their inputs and discussions.

Lastly, I would also like to thank my parents. Thank you for your support.

Abstract

Autism spectrum disorder (ASD) is a range of developmental brain disorders characterized by poor nonverbal communication skills, impaired behaviour and social interaction, and a limited diversity of social activities and interests. As the most common pervasive developmental disorder, it affects people of all economic backgrounds and races. Although the cause of ASD is still controversial, many studies indicated that genetic factors play an important role in the ASD pathology. It has been shown that there is a remarkable genetic heterogeneity among ASD cases and thousands of genes may be associated with this disorder.

The human brain can be separated into different regions in terms of their functions, and these regions are composed of distinct cell types. It has been shown that ASD risk genes are differentially expressed across different brain regions and at different embryonic stages. Some studies also revealed that the mutation of ASD risk genes may affect specific cell types more strongly than others. However, the extent to which ASD risk genes can converge on distinct cell types during human brain development remains unclear.

Recently developed single-cell RNA sequencing (scRNA-seq) dramatically advanced our knowledge of the cellular taxonomy of the brain and allowed us to map ASD risk genes or genomic loci onto specific brain cell types. In the present study, I aimed to uncover essential cell types underlying the development of ASD during human prefrontal cortex development by re-analysing a set of published scRNA-seq datasets. I mainly focused upon two sets of candidate ASD risk genes from the SFARI database, including 86 high confidence ASD risk genes (monogenic mutations in ASD) and 30 genes at the 16p11.2 locus (CNV in ASD). We found that distinct sets of ASD risk genes are enriched in neural progenitor cells, excitatory neurons, interneurons and glia cells. Such enrichments were due to cell subtypes within these major cell types having significant differences in ASD risk gene expression. Cell-type based gene network analysis further demonstrated that common signalling pathways and biological processes converged on cell types with enriched

expression patterns of ASD risk genes. Through comparative analysis, I further identified conserved and distinct expression patterns of ASD risk genes between human and mouse during brain development.

Taken together, this study provides important new insights into the cell type-specific molecular pathology of the ASD. The findings from this study also highlight the conserved and distinct functions of ASD risk genes implicated in the normal brain development between human and mouse.

Lay summary

Autism spectrum disorder (ASD) is a class of neurodevelopmental disorders featured by a remarkable genetic heterogeneity and thousands of different gene mutations may contribute to this disorder. The extent to which such genetic heterogeneity can converge on distinct cell types during the brain development remains unclear. In this study, I explored cell-type specific expression patterns of ASD risk genes in the human developing cortex and uncovered their functional importance in different aspects of human cortical development.

Recently it become possible to measure the gene expression in thousands of cells in developing brain tissues using a technology called single cell RNA sequencing. This has made it possible to study the expression of ASD risk genes in individual cells during brain development and identify cells that express large number of these risk genes. These cells maybe vulnerable to gene mutations causing ASD.

In the present study, I aimed to identify essential cell types associated with the ASD during the human brain development by re-analysing sets of published single-cell RNA sequencing data. I mainly focused on two sets of candidate ASD risk genes from the SFARI database, including monogenic mutations and copy number variant (CNV) mutation at the 16p11.2 locus in ASD cases. I found that distinct sets of candidate ASD risk genes are enriched in different cell types. Such enrichments are due to cell subtypes within these major cell types having significantly higher numbers of enriched ASD risk genes than others. Cell-type based gene network analysis further demonstrated that the common singling pathways and biological processes converged on cell types with enriched expression patterns of ASD risk genes. Through comparative analysis, we further identified conserved and distinct expression patterns of ASD risk genes between human and mouse during brain development. Collectively, these findings reveal vulnerable cell types underlying autism spectrum disorders in the developing human cortex.

List of abbreviations

AORUC –Area under an ROC Curve

ASD – Autism spectrum disorder

BP – Biological process

CC – Cellular component

CGE – Caudal ganglionic eminence

CNV – Copy number variation

CP – Cortical plate

CPM – Counts per million

DEG – Differentially expressed genes

DL – Deeper Layer

E – Embryonic day

ExNs – Excitatory neurons

FACS – Fluorescence-activated cell sorting

FDR – False discovery rate

GABA – Gamma-aminobutyric acid

GE – Ganglionic eminence

GO – Gene ontology

GW – Gestational week

GWAS – Genome-wide association study

INs – Interneurons

IPC – Intermediate progenitor cells

IZ – Intermediate zone

L – Layer

LGE – Lateral ganglionic eminence

MF – Molecular Function

MGE – Medial ganglionic eminence

MST – Minimum spanning tree

MZ – Marginal zone

NPC – Neural progenitor cell

OPCs – Oligodendrocyte precursor cells

oRGs – Outer radial glia cells

oSVZ – Outer subventricular zone

P – Postnatal day

PCA – Principal component analysis

PFC – Prefrontal cortex

PP – Preplate

Pvalb – Parvalbumin

QC –Quality check

RGC – Radial glial cell

RPKM – Reads per kilobase of transcript per million mapped reads

scRNA-seq – single-cell mRNA sequencing

SFARI – Simons Foundation Autism Research Initiative

SNP – Single nucleotide polymorphism

SP – Subplate

SST – Somatostatin

SVZ – Subventricular zone

t-SNE – T-distributed stochastic neighbour embedding

UP – Upper layer

vRG – Ventral radial glial cells

W – Developmental window

VIP – Vasoactive intestinal peptide

VZ – Ventricular zone

Table of Contents

Figure Index

Table Index

Chapter 1: General introduction

1.1 Genetic landscape of autism

Autism spectrum disorder (ASD) is a range of developmental brain disorders characterized by poor nonverbal communication skills, impaired behaviour and social interaction, and a limited diversity of social activities and interests (Constantino and Charman, 2016). As the most common pervasive developmental disorder, it affects people of all economic backgrounds and races. In the United Kingdom, more than one percent of people have ASD and close to four-fifths of the patients are male (Chung *et al.*, 2012). Although there is no known specific cause of ASD, some studies indicate that genetic factors play an important role in neurodevelopmental disorders, such as ASD. In 2003, based on the result of genome-wide association studies (GWAS) from three thousand families with autism, an evolving database was developed by Simons Foundation Autism Research Initiative (SFARI; http://sfari.org) (Banerjee-Basu and Packer, 2010). In this database, a list of genes was summarized whose mutation can contributes to ASD. Different types of genetic heterogeneity, such as single-gene mutations and single nucleotide polymorphisms (SNP) are associated with ASD. During the last decades, the SFARI Gene Scoring Advisory Panel grouped these genes based on a “Gene Scoring Module”, which was established based on the published literature on the genetics of autism (Table 1). The Gene Scoring Module offers critical evaluation of the strength of the evidence for each gene’s association with ASD, and helps research community establish criteria for assessing the strength of the evidence linking candidate genes to ASD. In detail, the genes included in single-gene mutations (also called as monogenic genes) are separated into six categories based on the strength of the evidence linking candidate genes to ASD: high confidence, strong candidate, suggestive evidence, minimal evidence, hypothesized but untested, and evidence does

not support a role. The evidence of ASD risk genes not only includes research from human genetics study, but also functional studies of these risk genes in both human and gene-knockout mice. Besides these monogenetic genes, the copy number variance (CNV) of genetic loci (CNV genes), either deletions or duplications, are also linked to this disorder (Figure 1). These genes, usually called “SFARI genes”, or “ASD risk genes”, are highly referenced by the autism research community and have been studied in a large number of projects. In this thesis, the analysis of gene expression will focus on the 86 highest rank candidate ASD risk genes (“Category 1” and “Category 2” in Table 1), as these genes are significant statistically in genome-wide studies between cases and controls. Genes on *16p11.2* locus will also be included as both duplication and deletion of genes on this locus has been linked to significantly increased incidence of ASD (Lin *et al.*, 2015).

Table 1: Categorisation of ASD monogenic risk genes.

This is based on the critical evaluation of the strength of the evidence for each gene's association with autism. These scoring criteria are offered by SFARI website (https://gene.sfari.org/database/gene-scoring/).

Category	Criteria
Category 1 (High confidence) Category 2 (Strong candidate)	A rigorous statistical comparison between cases and controls, yielding genome-wide statistical significance, with independent replication, to be the strongest possible evidence for a gene. These criteria were relaxed slightly for category 2.
Category 3 (Suggestive evidence) Category 4 (Minimal evidence)	The literature is replete with relatively small studies of candidate genes, using either common or rare variant approaches, which do not reach the criteria set out for categories 1 and 2. Genes that had two such lines of supporting evidence were placed in category 3, and those with one line of evidence were placed in category 4.
Category 5 (Hypothesized but untested) Category 6 (Evidence does not support a role)	The list of genes in SFARI Gene is inclusive, and as such there are genes that have been implicated solely by evidence in model organisms or other evidence of a marginal nature. These genes were placed in category 5, as they have not yet been rigorously tested in a human cohort. Category 6 is for those genes that have been tested in a human cohort, but the weight of the evidence argues against a role in autism.

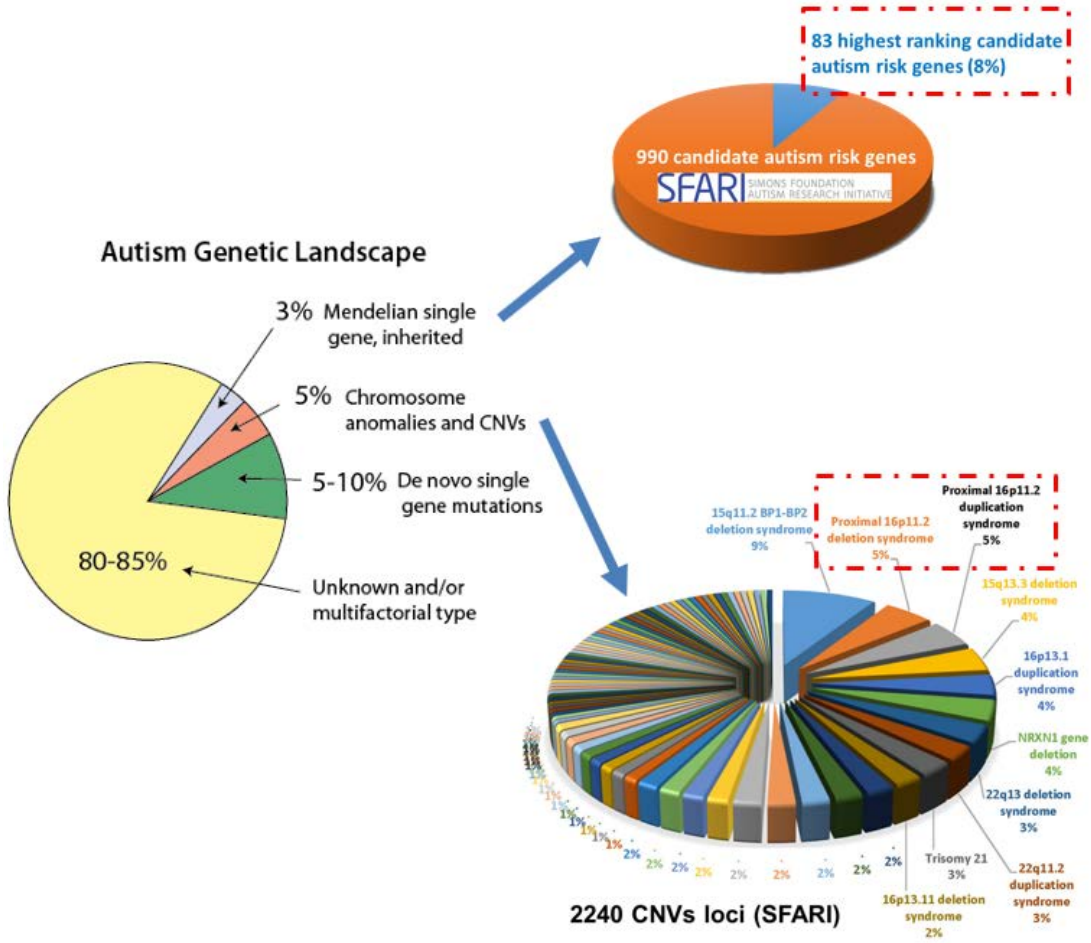


Figure 1: The genetic landscape of ASD identifies which known genes and types of variations are responsible for the disorder.

The highest-ranking candidate ASD risk genes, which labelled as “high confidence” and “strong candidate” in SFARI database, and genes on *16p11.2* locus are highlighted in red box. Figures were modified from https://corticalchauvinism.com/2017/03/06/genetic-counseling-in-autism/.

A review highlighted that embryonic neurogenesis maybe a potentially important locus of pathology in ASD during the human brain development (Packer, 2016). In the review, the authors summarized that many ASD risk genes may be involved in neural progenitor proliferation and migration, and the converge on events that precede synaptogenesis, including the proliferation of neural progenitor cells and the migration of neurons to the appropriate layers of the developing neocortex. In some instances, this pathology may be driven by alterations in chromatin biology and canonical Wnt signaling, which in turn affect fundamental cellular processes such as cell-cycle length and cell migration. They also reviewed that some ASD risk genes regulate specific biological processes at early fetal stages across developmental trajectories, such as progenitor proliferation and neural migration. Further, these ASD genes were highly co-expressed in gene-gene networks that implicate distinct biological functions during human cortical development, such as early transcriptional regulation and synaptic development. At laminar-specific level, the expression levels of these genes are enriched in glutamatergic neurons and superficial cortical layers.

Through application of mRNA sequencing, the fields understanding of autism disease mechanisms through genetics has proliferated in recent years. However, efforts to analyse the spatiotemporal dynamics of ASD risk genes expression across different cell types, especially the possible biological function of ASD risk genes in each cell type at different developmental stages, is lacking. It’s important to consider that the diversity of cortical neurons has typically been defined based on criteria of morphology, electrophysiology, ontology, and the expression of a few transcripts and proteins. And several molecular or cellular mechanisms are leading to the diversity of cortical neurons during the development of fetal brain, including neurogenesis, differentiation, migration and synaptic function. But these mechanisms are not entirely distinct, and more work is needed to refine these mechanisms to functional pathways. The same genes or molecular pathways contribute to several of these mechanisms at different points during development, and it is not totally clear how early developmental dysfunction relates to ASD.

A set of molecular or cellular mechanisms are depicted from early fetal to neonatal stages with the progress of cortical development. The composition of cell types changes across these laminae, and each of those cell types expresses a distinct set of genes and plays a unique and essential role in the development and functions of the fetal brain, as well as effect on the developmental disorders. For example, Willsey et al. observed candidate cell populations likely to be disrupted with selective vulnerability during early development (Willsey *et al.*, 2013). The authors micro-dissected different layers of the mouse cerebral neocortex and investigated the gene expression levels of layer-specific tissue. Based on the knowledge of marker genes of different cell types, the authors measured cell type proportions broadly using cell type-specific gene expression references and revealed that most of ASD risk genes were highly expressed in the deep layer cortical glutamatergic neurons.

Furthermore, the same set of genes may play different roles or show different expression patterns at different points during development. For example, based on the co-expression analysis of gene expression profiling from bulk mRNA sequencing, the scientists found that the genes on *16p11.2* locus encoded some co-expressed interacting protein pairs at different developmental stages in different human brain regions (Lin *et al.*, 2015). In detail, KCTD13 co-expressed with other proteins that were encoded by the genes on *16p11.2* locus. During late mid-fetal development, the bioinformatics analysis revealed that these proteins played roles in DNA replication and synthesis in the prefrontal and motor-sensory cortex. In the parietal, temporal, and occipital cortex, KCTD13 co-expressed network not only takes part in DNA replication and synthesis, but also regulates the formation of E3 ubiquitin ligase complexes. A higher number of pairs between the proteins encoded by these genes were found to be co-expressed and interacting in the cortex during late fetal development. In the early fetal developmental stages, only around 10% of *16p11.2* proteins co-expressed, and this fraction increased to more than 30% during late mid-fetal developmental stages.

In the previous works, such as Willsey's and Lin's studies, they dealt with bulk RNA sequencing data in complex brain tissues, and the gene expression levels were measured based on the average expression levels of genes across different cell types. However, it is hard to characterize the cell type compositions exactly from bulk RNA sequencing data. Meanwhile, considerable genetic and cellular heterogeneity has complicated efforts to establish the biological foundations of ASD, and these results require consideration of the diversity of cell types in fetal brain. A big change is currently underway in the genomics of ASD. Transcriptional profiling of individual cells has emerged as an essential tool for characterizing cellular diversity. The recent development of high-throughput single-cell mRNA sequencing (scRNA-seq) techniques allows us to profile gene expression at the single-cell level easily, and has led to the systematic discovery of detailed biological effects of ASD risk genes across different cell types in course of brain development.

1.2 Diversity of cell types during cortical development

Understanding cellular diversity in the brain has been a long-standing question in neuroscience. The whole development of human embryonic neocortex can be divided into five stages based on the formation of cortical structure (Figure 2). At the beginning of human cortex early development, around the first four weeks, there is only one zone called ventricular zone (VZ), and a preplate (PP) covers this zone. With development processing, the PP divides into two different laminae called marginal zone (MZ) and subplate (SP) around the 4 gestational weeks (GW) old. At the same time, the VZ will extend to more layers, such as subplate zone/intermediate zone (SP/IZ) and cortical plate (CP). With time going, the SP region change to a compartment called the SP/IZ and a middle layer called the CP (Olson, 2014). In the GW8, subventricular zone (SVZ) is formed in SP/IZ region and this zone will be further differentiated

into two distinct germinal compartments that are known as the inner SVZ (ISVZ) and a massively expanded outer region (OSVZ) after the first trimester (Lewitus, Kelava and Huttner, 2013; Dehay, Kennedy and Kosik, 2015). Further proliferation of neuronal cells can induce the formation of the outer fibre layer (OFL). As the depth of SVZ is growing, an inner fibre (IFL) become large to split the ISVZ and OSVZ.

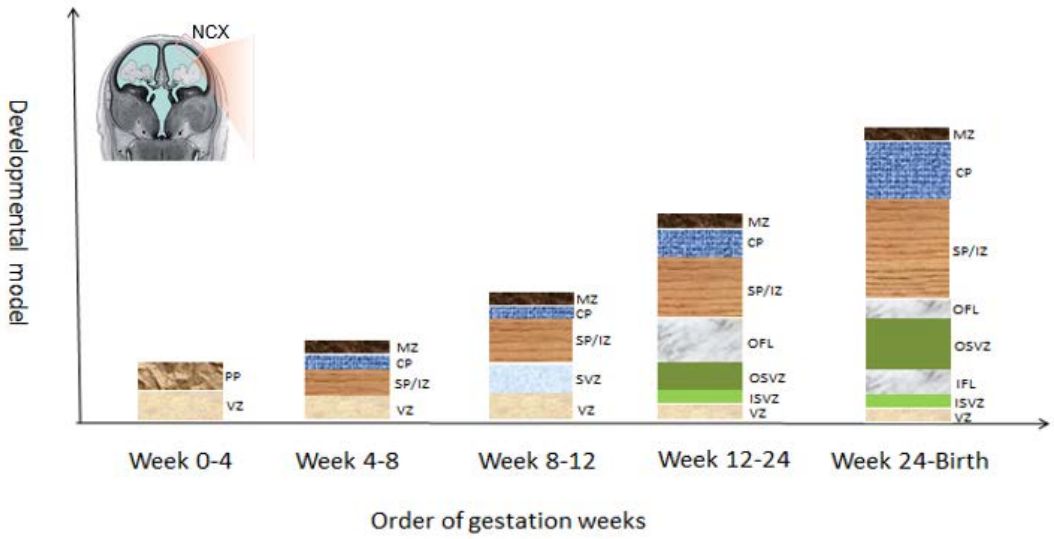


Figure 2: Schematic representation of the early development of human embryonic cortex.

During the processes above, cells are differentiating into different cell types to make the human cortex perfect both in structure and function. At the beginning, the self-renewal neural stem cells (NSCs) in the VZ can develop into radial glia (RG) cells and generate new intermediate progenitor cells. Then in the MZ, a superficial layer of pioneer neuron expands from PP. Majority of these pioneer neurons are called as Cajal-Retzius (C-R) cells. These cells can secrete a chemical compound which let neurons migrate to CP. SP cells are produced at the same time as when C-R cells are generated at the MZ (Hansen *et al.*, 2010). The cell types between IZ and SVZ are similar, and only some fibres

can be found in SP. These parts are usually called as SP/IZ. The process in human is different with mouse as SVZ is developed after SP/IZ formation in human (Reith, 1996). Further, SVZ can be separated into two parts, ISVZ and OSVZ. Outer radial glia progenitor cells (oRGs) and intermediate progenitor cell (IPCs) are involved in ISVZ and OSVZ, respectively. OFL and IFL are the floors which separate the ISVZ and OSVZ, respectively. We can regard ISVZ, IFL, OSVZ and OFL as a whole SVZ region in corticogenesis. All kinds of NSCs and RGs are often referred together as neural progenitor cells (NPCs). Progenitor cells and IPCs in both VZ and SVZ are responsible for the generation of cortical excitatory neurons (ExNs). ExNs in the human cerebral cortex are generated in a limited period of development, from about GW5 to about GW20 (Costa and Müller, 2015). The broadest classification of cortical neurons splits them in two large groups: the ExNs and the inhibitory interneurons (INs). The origin of INs are not cortex. After interneurons are born in the ganglionic eminences (GEs), they begin to tangentially migrate towards the developing cortex. Immature cortical interneurons travel long distances before reaching the cortical place, where they shift to radial migration to reach their final destination along the cortical layers (Anderson, 2001; Bartolini, Ciceri and Marín, 2013). In mice, migration of cortical interneurons begins at embryonic day (E)12.5 and is completed by birth, when integration into circuits begins (Anderson, 2001). However, the detailed time points about the differentiation and migration of cortical interneurons in human is not get clear.

Besides NPCs, ExNs and INs, there are other three kinds of cell types that exist in human cortex. Microglia cells are a type of glial cell located throughout the brain and generated between GW4 and GW24 (Menassa and Gomez-Nicola, 2018). Astrocytes are the most abundant cell type in the adult human brain and have many important physiological functions, and they are largely produced during the early postnatal stages (Reemst *et al.*, 2016). Oligodendrocyte progenitor cells (OPCs) proliferate and migrate away from ventricular germinal zones of the embryonic neural tube into developing gray and white matter before differentiating into oligodendrocytes (Bergles and

Richardson, 2016). The three cell classes above are not introduced in detai since they are not discussed in this thesis.

These six classes of cells (NPCs, ExNs, INs, Microglia, Astrocytes and OPCs) are regarded as cardinal cell classes in the developing human cortex. Over the course of development, these cell classes are generated in specific developmental stages and build different kinds of functional circuits across different laminae. The gene expression across these cell classes are variable to maintain cell class-specific signatures. The expression levels of marker genes across different cell classes are distinct and the gene expression levels in the same cell type are also not static throughout their lifetime. During cortical development, cells undergo a variety of molecular and genetic changes. Several molecular or cellular mechanisms lead to the development of human fetal brain, including neurogenesis, differentiation, migration and synaptic function. Most of these mechanisms are individually quite widely affect the different processes, and more works are needed to refine these mechanisms to functional pathways. However, these mechanisms are not entirely distinct. The same genes or molecular pathways contribute to several of these processes at different points during development, and it is not totally clear how early developmental dysfunction relates to ASD.

A set of molecular or cellular mechanisms are depicted from early fetal to neonatal stages with the progress of cortical lamination (Figure 3). The numbers on the timeline indicate the molecular pathways related with lamination important at the stages of development. The composition of cardinal cell classes is changing across these laminae and like the study about molecular or cellular mechanisms, each of those cell classes expresses a distinct set of genes and plays a unique and essential role in the development and functions of the fetal brain. Furthermore, the same set of genes maybe play different roles to several of these cell types and/or laminae at different points during development. Based on Kang's work and Willsey's study, for human fetal cortical development, the fourty GWs can be separated into four ages: early fetal (GW6-15), mid fetal (GW15-21), mid-late fetal (GW21-26) and

late fetal (GW26-40). The human brain development is characterized by a chain of cellular events and functional milestones including the expansion of neural stem cell/progenitor cell pool, neurogenesis, cell type differentiation, neuronal migration, emergence of and disappearance of transient cellular compartments (VZ, SVZ, MZ, SP and CP), axonogenesis and dendrite growth, gliogenesis and neuronal circuit formation. In humans, these cellular changes and underlying neurodevelopmental mechanisms are best understood in the cerebral neocortex.

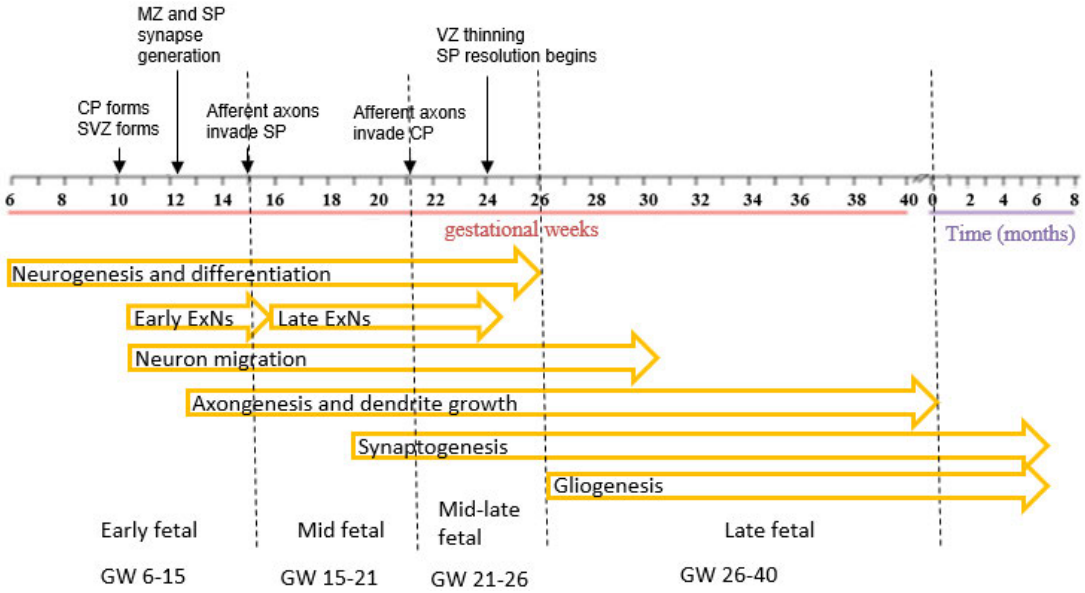


Figure 3: Timeline of key cellular processes and functional milestones in the human developing brain.

The figure provides an overview of some key cellular processes in the human developing cortex and functional milestones. The top panel shows the key functional milestones and their timing during human brain development. The second panel provides a timeline of human brain development, and age in postconceptional weeks, and postnatal months. The bottom panel details the time span and sequence of key cellular processes in the developing brain. Bars indicate the peak developmental period in which each feature is acquired. Figure is modified based on the figure in a review article about the understanding of autism genetics (De La Torre-Ubieta *et al.*, 2016).

In summary, a set of molecular or cellular mechanisms are depicted from early fetal to neonatal stages with the progress of cortical lamination. And the composition of cardinal cell classes is changing across these laminae and like the study about molecular or cellular mechanisms, each of those cell classes expresses a distinct set of genes and plays a unique and essential role in the development and functions of the fetal brain. Furthermore, the same set of genes maybe play different roles to several of these cell types and/or laminae at different points during development (De La Torre-Ubieta et al., 2016).

1.2.1 Neural progenitor cells in the developing cortex

As previously described, the human cortex develops from two principal germinal zones, the VZ and the SVZ. And notably in humans, but not rodents, an inner SVZ (ISVZ) and an outer (OSVZ) can be distinguished (Smart, 2002; Dehay, Kennedy and Kosik, 2015). Correspondingly, the VZ and SVZ harbour the cell bodies of three principal classes of NPCs, called vRGs, oRGs and IPCs. The locations of these cell types are distinct during human cortical development (Figure 4). Studies dissecting the progress between NPCs proliferation and differentiation have demonstrated that most of vRGs are located at VZ and a few of vRGs are located at ISVZ. oRGs are only located at OSVZ, and IPCs can be found at both VZ/ISVZ and OSVZ.

To further characterize this diversity, scientists compared gene expression signatures across the three cell types (Figure 4B). The canonical progenitor markers *SOX2* and *PAX6* were expressed at similar levels in all germinal zone regions, as well as all progenitor cell types. *HES1* and *VIM* were expressed at similar levels in both oRGs and vRGs, but not expressed in IPCs. To test the distinction between vRG and oRG cells, scientists identified a set of genes differentially expressed between the VZ and the OSVZ regions in the GW23 BrainSpan Atlas data. They revealed *CRYAB* and *ANXA1* showed signal only in the VZ, where vRG cells enriched. *HOPX* and *FAM107A* represent markers

for human developing oRGs since these genes were enriched in OSVZ region at mid-gestation. High *EOMES* and *HES6* expression was detected in IPCs but not in vRG or oRG cells.

These studies validated the diversity of progenitor cell types during human cortical development, and strongly suggest that the transcriptomic profile across these cell types could be different to fit the function of these cells. It will be interesting to look at the expression pattern of ASD risk genes

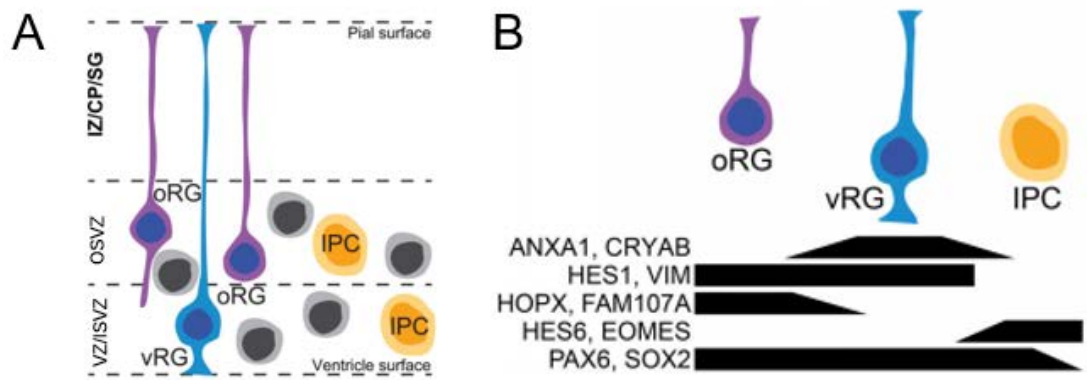


Figure 4: Schematic representation of the development of human progenitor cells.

(A) The diversity of progenitor cell types during development. (B) Summary of marker genes of oRG, vRG and IPCs cell types identified from previous studies. Figures are cited from a transcriptomic study of human radial glial diversity (Thomsen *et al.*, 2016).

1.2.2 Excitatory neurons in the developing cortex

During development, ExNs acquire their identity and regional position depending on the location of their progenitors, while their layer position is defined by the time of their birth, in an “inside-out” sequence (Kriegstein, Noctor and Martínez-Cerdeño, 2006; Suzuki and Vanderhaeghen, 2015) (Figure 5). The generation of the ExNs is radial migrated from VZ to CP both in human and rodents. In the dorsal telencephalon, the neuroepithelial cells (NE) first expand by symmetric proliferative divisions during early development and then convert to RGs to initiate neuron production, either directly or through transit progenitors in the SVZ, including oRGs and IPCs. This will result in the sequential generation of early deep layer (DL) and late upper layer (UL) neurons.

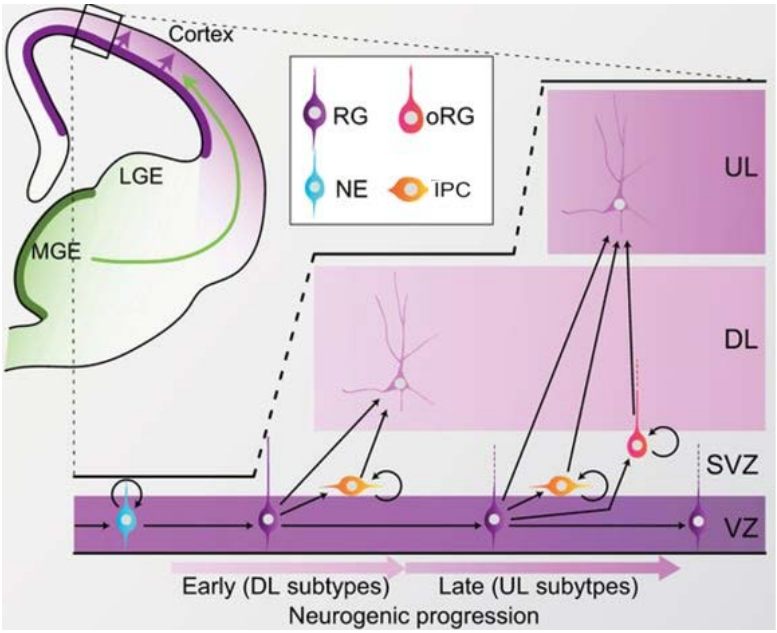


Figure 5: Radial migration of excitatory neurons at embryonic stages.

The different colors indicate different types of cells. Blue, NE; Purple, RGs; Orange, IPCs; Red, oRGs. During neurogenic progression, the early generated ExNs are migrated from VZ to deep layer, and the late generated ExNs are migrated from VZ to upper layer. The purple direction within cortex stands for the radial migration, and the green direction from MGE to cortex stands for the tangential migration (top left). The tangential migration will be explained in the next INs section (Suzuki *et al.*, 2015). This schematic representation of the development of ExNs do not specific for human or rodent.

These ExNs subtypes show distinct spatial organization that can be characterised by specific molecular and functional properties. In this thesis, we focus on the transcriptomic profiles of ExNs subtypes in corticogenesis. In the 'inside-out' fashion, the oldest neurons (early generated) tend to be located near the pial surface. Successive waves of newly generated neurons (late generated) migrate past the existing early born neurons, and migrate to the ventricular surface of cerebral cortex, creating cortical layers (L) 2-6 (Figure 6 left). Figure is cited from a review of human ExNs development (Gao *et al.*, 2013; Van den Ameele *et al.*, 2014).

Neurons from different layers are produced at different developmental time points, and the adult cerebral cortex consists of six layers. In an early transcriptomic study targeting neocortical layers, J. G. Chen et al. (2005) microdissected upper layers (L2–L4) and deep layers (L5 and L6) from early postnatal mice (P7) for microarray analysis. They found that transcription factor Zfp312 is selectively expressed by L4-L5 ExNs and their progenitor cells (Chen *et al.*, 2005). In a more comprehensive study, Fertuzinhos et al. (2014) employed mRNA sequencing to gain insights into transcriptional events involved in laminar development. They microdissected of infragranular layers (IgL, deep layers, L5-L6), granular layer (L4), and supragranular layers (superficial layers, upper layers, L2-3) from a seires of postnatal mice (P4, P6, P8, P10, P14, and P180) (Fertuzinhos *et al.*, 2014). They identified 662 protein-coding genes significant differentially expressed across layers and 1,321 protein-coding genes significant differentially expressed across ages (false discovery rate (FDR) <0.01). In a more recent study, He et al. (2017) characterized the transcriptome of the cortical layers (L2-L6) of adult human prefrontal cortex. Based on the high-throughput sequencing, they characterized 2,320 human layer markers (FDR < 0.05). The high-precision study of RNA in situ hybridization from Allen Human Brain Atlas illustrated the layer-specific expression pattern of selected markers in the adult human temporal cortex (Figure 6, right).

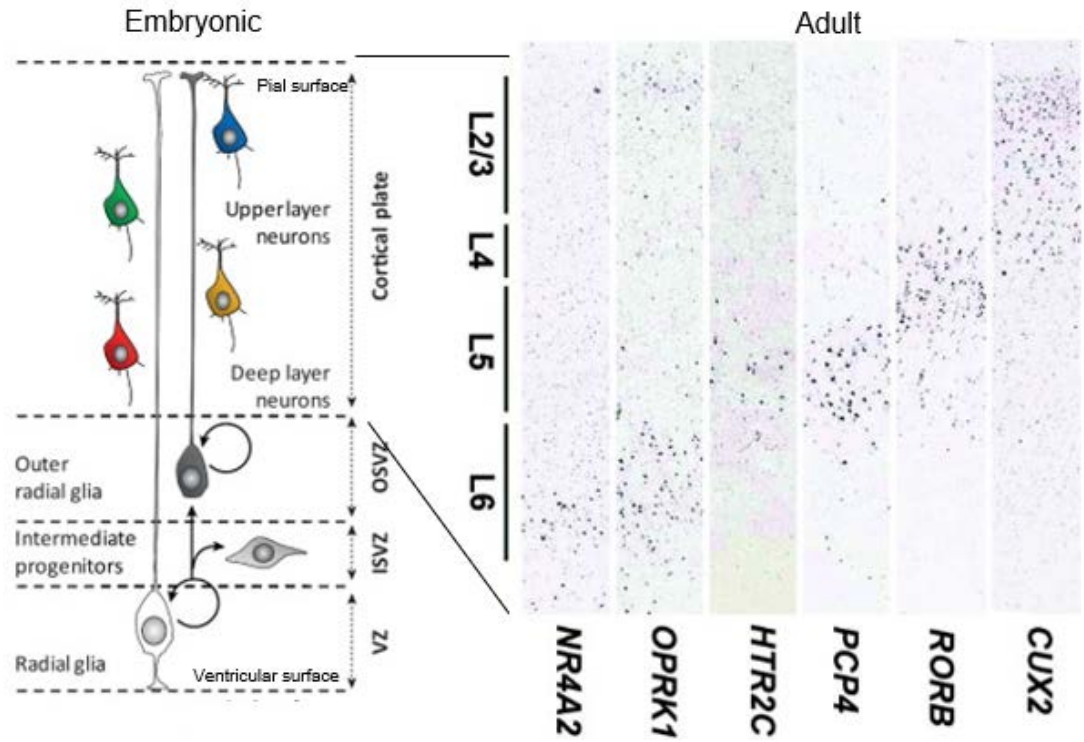


Figure 6: Excitatory neuronal subtypes showing distinct spatial organization.

The ExNs are layered according to birth date. The colored cells at the left is meant to indicate the relative spatial organization of DL and UL neurons during human cortical development. RNA in situ hybridization from Allen Human Brain Atlas at the right shows layer-specific expression of selected markers in the adult temporal cortex. The dark grey points indicate cells positive for *NR4A2* (Layer 6), *OPRK* (Layer 6), *HTR2C* (Layer 5), *PCP4* (Layer 5), *RORB* (Layer 4) and *CUX2* (Layer 2/3) in human adult cortex. Figure at left is cited from a transcriptomic study of human developing cortex (Lake *et al.*, 2016). Figure at right is cited from a in-situ hybridization study of human adult cortex (Zeng *et al.*, 2012).

1.2.3 Interneurons in the developing cortex

The broadest classification of cortical neurons splits them in two large groups: the ExNs and the INs. The balance between the two neuronal cell types is indispensable for the normal fucntion of neuronal circuits (Rossignol, 2011; Lewis *et al.*, 2012; Lin and Sibille, 2013; Volk *et al.*, 2015). In adult cortex, across difference species and brain regions, ExNs are glutamatergic, myelinated, long-projecting cells that correspond to approximately 80% of all cortical neurons. In contrast to ExNs, INs are GABAergic, inhibitory, local-projecting cells that correspond to approximately 20% of all cortical neurons (Hendry *et al.*, 1987). But recently, scientists find that some GABAergic neurons in the cortex and hippocampus are long-projecting neurons (Melzer *et al.*, 2017).

In contrast to radial migration of cortical ExNs, cortical INs are migrated tangentially from three different interneuron progenitor regions with the expression of different transcription factors. Two regions of ganglionic eminences (GEs), including medial and caudal ganglionic eminences (MGE and CGE, respectively), and the preoptic area (POA) are the origin of cortical INs (Gelman and Marín, 2010).

Cortical INs are remarkably diverse and the transcriptome analysis show that the MGE, CGE and preoptic region generate different classes of mouse cortical INs (Gelman and Marín, 2010). In this thesis, we will focus on the diversity of MGE and CGE- derived cortical INs. It was well known that some transcription factors, such as *Dlx1/2*, *CoupTF1/2*, *Gsx1/2*, *Arx* and *Npas1/3*, are involved in regulating both MGE- and CGE-derived interneuron fates due to their broader expression in the MGE and CGE progenitors (Cobos *et al.*, 2005; Colasante *et al.*, 2008; Kanatani *et al.*, 2008; Lodato *et al.*, 2011; Cai *et al.*, 2013; Stanco *et al.*, 2014) (Figure 7). Expression of *Nkx2.1* in progenitor cells is only found in the MGE, but not CGE. *Lhx6* lies downstream of *Nkx2.1* to regulate MGE specification and neurogenesis (Sussel, 1999; Du *et al.*, 2008). *Sox6*, a transcription factor, is expressed in most MGE-derived cortical

interneurons. There are also a few of transcription factors that have been shown to specifically regulate CGE interneuron fate. For example, *Prox1*, a homeodomain transcription factor, is initially expressed in the SVZ within the ganglionic eminences, becomes restricted to CGE-derived cortical interneuron precursors and is selectively maintained within this population in the adult cortex (Rubin and Kessaris, 2013; Miyoshi *et al.*, 2015).

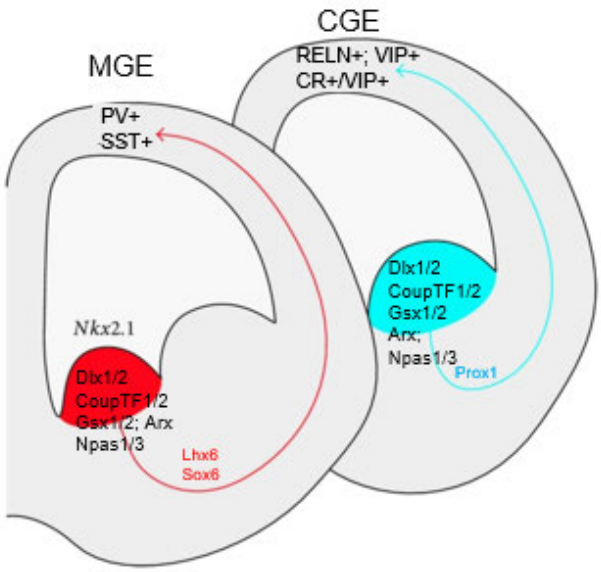


Figure 7: Schematic representation of genetic cascade during the generation of mouse cortical interneurons.

Red, MGE region; Blue, CGE region. The MGE generates parvalbumin-positive (PV+) interneurons, as well as somatostatin-positive (SST+) interneurons. The CGE generates reelin-positive (RELN+) interneurons, vasointestinal peptide-positive (VIP+) interneurons, and VIP and calretinin (CR) double-positive interneurons (Gelman and Marín, 2010).

Previous stuies indicated that the MGE and CGE regions generate different classes of cortical interneurons (Figure 7). The MGE-specifc transcript factors regulate the development of SST+ and PV+ cortical interneurons (Wichterle *et al.*, 2001; Xu *et al.*, 2004; Xu, Tam and Anderson, 2008; Gelman *et al.*, 2011), whereas the CGE produces VIP+ cortical interneruons, RELN+ cortical interneurons. and VIP+/CR+ cortical interneurons (Nery, Fishell and Corbin, 2002; Xu *et al.*, 2004; Butt *et al.*, 2005; Miyoshi *et al.*, 2015; Niquille *et al.*, 2018).

Recently, with mRNA sequencing of individual cells, scientists have analysed the transcriptional profiles of specific groups of interneurons with high precision. In one of the first studies of single cell transcriptomes, Tasic et al. (2018) collected 1,424 indivadul interneurons from the adult mouse visual cortex, clustered single cells based on their transcriptomes, and revealed 49 interneuron cell types, which were identified by the expression of marker genes (Figure 8) (Tasic *et al.*, 2018). *Sst*, *Pvalb* and *Vip* genes are identified as marker genes of SST+, PV+ and VIP+ interneruon, respectively.

They also identified new molecularly defined subpopulations of SST+, PV+ and VIP+ interneurons, suggesting the raised awareness on the molecular heterogeneity of cortical interneurons.

Cortical interneurons are remarkably diverse. There is a partial conservation of diversity between mouse and humans in cortical INs. For example, the major cell types of cortical INs (SST+, PV+ and VIP+) and the molecular mechanisms of migration are conserved between mouse and humans developing cortex (Hodge *et al.*, 2018). Molecular mechanisms that control the fate determination of MGE and CGE-derived interneurons in human have remained largely elusive, partially due to the lack of good molecular markers to specifically label this region and the lack of experimental materials.

Besides the variability of molecular markers and embryonic origins among cortical INs, the main features exposing such diversity include dendritic and axonal morphology, synaptic characteristics and firing properties (Ascoli *et al.*,

2008; Rudy *et al.*, 2011; Kepecs and Fishell, 2014). Howerver, this thesis will consider the diversity of interneurons based on the molecular markers since only the gene expression profiling data are accessed.

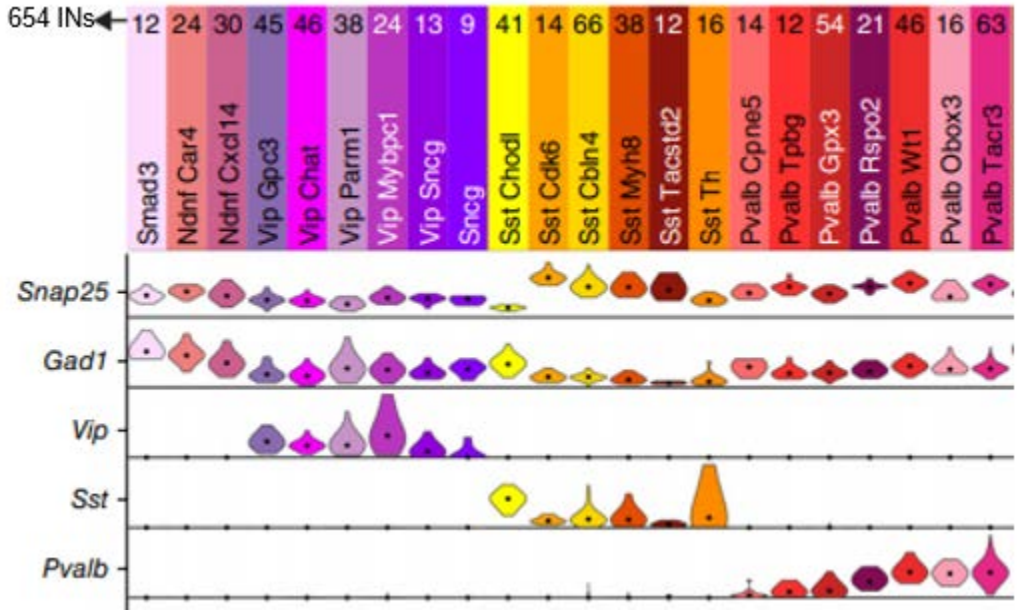


Figure 8: Remarkably diversity of cortical interneuron cell types and marker gene expression.

Each type is represented by a color bar with the name and number of cells representing that type. The violin plots represent distribution of marker gene expression for each cell type. Snap25 is marker gene of cortical neurons; Gad1 gene is marker gene of cortical INs; Sst, Pvalb and Vip genes are marker genes of SST+, PV+ and VIP+ interneurons, respectively. Figure is cited from a single cell transcriptomic study of mouse visual interneurons (Tasic *et al.*, 2018).

1.3 General introduction of single cell mRNA sequencing

Bulk mRNA sequencing studies have generated remarkable insights and resources on the development of the cerebral cortex (Silbereis *et al.*, 2016). But they only provide transcriptomic data at a population level. Compared to bulk mRNA sequencing, single cell mRNA sequencing (scRNA-seq), for example, the result in Figure 8, has the ability to characterize new or rare cell types in a tissue population.

By sequencing the transcriptomes of different cell types in a cell population, scRNA-seq is more sensitive, accurate and reproducible than the traditional bulk mRNA sequencing (Figure 9). For example, there are many different cell types in the development of human cortex. Based on the traditional bulk tissue RNA sequencing, we can only know the differences of gene expression in the whole region between different time points. While through scRNA-seq, we can classify individual cells into different cell types, then we can know not only the differences in the whole region, but also the differences between cell types in the same or different time points. Transcriptional profiling of scRNA-seq is useful to identify the diversity of cells in a tissue, as well as to analyse the different expressed genes among the cell types (Saliba *et al.*, 2014).

Since the transcriptome reflects the functional properties of a cell at a given time, single cell transcriptomics is perhaps the most comprehensive approach available for the classification of cellular diversity to date. Recently, it has become possible to measure the gene expression in thousands of cells in a tissue based on the two latest scRNA-seq platforms. The first one is Fluidigm C1. The C1 system allows cell capture, lysis, reverse transcription, and cell multiplexing occur in an integrated fluidic circuit chip. 96 cells can be collected per chip. C1 technology is usually combined with SMART-sequencing and generates full-length cDNA library. It can detect about 9,000 genes per cell. The second platform is 10X Genomics Chromium. It performs rapid droplet-based encapsulation of single cells, then lyse the cells in the droplets thereby

releasing cellular mRNAs and build cDNA library from 3′ poly A tails. This platform allow high throughput (possible up to 10,000 cells), and currently able to detect about 4,000 genes per cell. With the developing of technology, the new protocol allowing a total of 9,600 cells to be multiplexed and sequenced together (Svenson *et al.*, 2018).

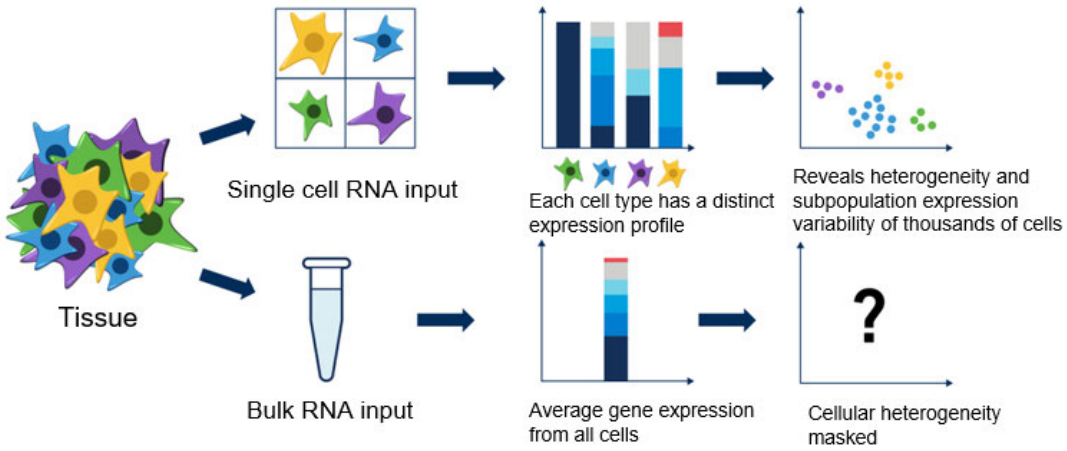


Figure 9: Brief comparison between bulk RNA sequencing and scRNA-seq.

scRNA-seq generates gene expression profiles at the resolution of individual cells. Based on the expression profile of individual cells, scientists can identify cellular diversity, and reveal gene expression pattern among cell types. The bulk mRNA sequencing can only assess transcriptome at a tissue level, and calculate the average expression level of multuple cell types in the tissue. The average gene expression pattern will miss the cell-to-cell variability. Figure is cited from https://www.biocompare.com/Bench-Tips/345311-Single-Cell-Set-Up-Sample-Preparation-Tips/.

1.4 Aim of this thesis

Autism spectrum disorder (ASD) is a class of neurodevelopmental disorders featured by a remarkable genetic heterogeneity, with thousands of different genes may contribute to this disorder. The extent to which such genetic heterogeneity can converge on distinct cell types during the brain development remains unclear. Recently developed single-cell RNA sequencing dramatically advanced our knowledge of the cellular taxonomy of the brain and allow us to evaluate whether the ASD risk genes or genomic loci map onto specific brain cell types. In this thesis, we aim to identity essential cell types associated with the ASD during human brain development by reanalyzing sets of published single-cell RNA sequencing data. Two sets of candidate ASD risk genes from the SFARI database will be used, including 86 high confidence ASD risk genes (monogenic mutations in ASD) and 30 genes at the *16p11.2* locus (CNV genes in ASD). We plan to illuatrate if any distinct sets of candidate ASD risk genes are enriched in any neural cell types during human corical development. We also try to reveal the expression pattern of ASD risk genes within mouse developing INs of the MGE, CGE and cortex.

Chapter 2: Materials and Methodology

2.1 Materials

Recently, scRNA-seq method have been applied to the developing human and mouse cortex, which has led to remarkable progress in understanding the molecular signatures that define cortical cell types and developmental progress within cortex. Some important questions about the developmental processes in cortical development have been solved successfully from the scRNA-seq studies. For example, scientists find out the molecular characteristics that establishes the oRG identity during human cortical development (Pollen *et al.*, 2015; Nowakowski *et al.*, 2016), as well as identified molecular characteristics that guide the sequence of neuronal differentiation events in the mouse and human cortex at early developmental stages (Kageyama *et al.*, 2018; Loo *et al.*, 2019). These studies provided a rich data resource for further studies. Here we selected the data from six of these publications, and we re-analysed the data from these publications to explore the gene expression patterns of ASD risk genes across different cortical cell types in both human and mouse (Table 2). The details of each dataset are described in the Result chapters.

Table 2: Table summarizing the datasets used in this thesis.

PFC: prefrontal cortex; FC: frontal cortex; VC: visual cortex; AC: anterior cortex; V1: primary visual cortex; MGE: medial ganglionic eminence; CGE: caudal ganglionic eminence. GW: gestational week; E12.5/E14.5/E18: embryonic day 12.5/14.5/18 of development.

Datasets	Method	Ages	# of cells	Types	Species	Brain Region
Zhong *et al. Nature. 2018*	Smart-Seq2	GW8 to GW26	2306	Fetal	Human	PFC
Nowakowski *et al. Science. 2018*	C1	GW6 to GW37	4261	Fetal	Human	PFC, V1 and MGE
Lake *et al. Science. 2018*	C1	51 years old	2478	Adult	Human	PFC,FC,VC and AC
Mi *et al. Science. 2018*	C1	E12.5 and E14.5	2003	Fetal	Mouse	MGE and CGE
Mayer *et al. Nature. 2018*	10X	E18	8382	Fetal	Mouse	Cortex
Tasic *et al. Nature Neuroscience. 2018*	SMARTer	8 weeks old	766	Adult	Mouse	Visual Cortex

2.2 Bioinformatics analysis of scRNA-seq data

This section was concerned with the computational analysis of the data obtained from scRNA-seq studies (Figure 10). The first steps, called data pre-processing, were general for any high throughput sequencing data. Starting from sequencing reads, these steps contain the steps required for quality control to avoid adapter contamination and higher error rates in reads boundary (read QC, Alignment and Mapping QC), remove problematic cells (Cell QC), and normalization of cell-specific biases (Normalization). Since we

used the published datasets in this thesis, these steps were not conducted in our analysis except normalization.

Later steps required a mix of existing bulk RNA-seq analysis methods and novel methods to address the technical difference of scRNA-seq. All analysis was conducted using software packages from the open-source Bioconductor project (release 3.5) (Huber *et al.*, 2015). Starting from a normalized gene expression matrix, the application of different steps in the workflow will be demonstrated on different aims involving identification of cell types, calculation of differential expressed genes, dynamic expression of genes along developmental trajectories, the assignment based on the correlation between embryonic and adult cells, and the gene ontology (GO) term analysis of enriched genes in each cell type. The details of each step were described below.

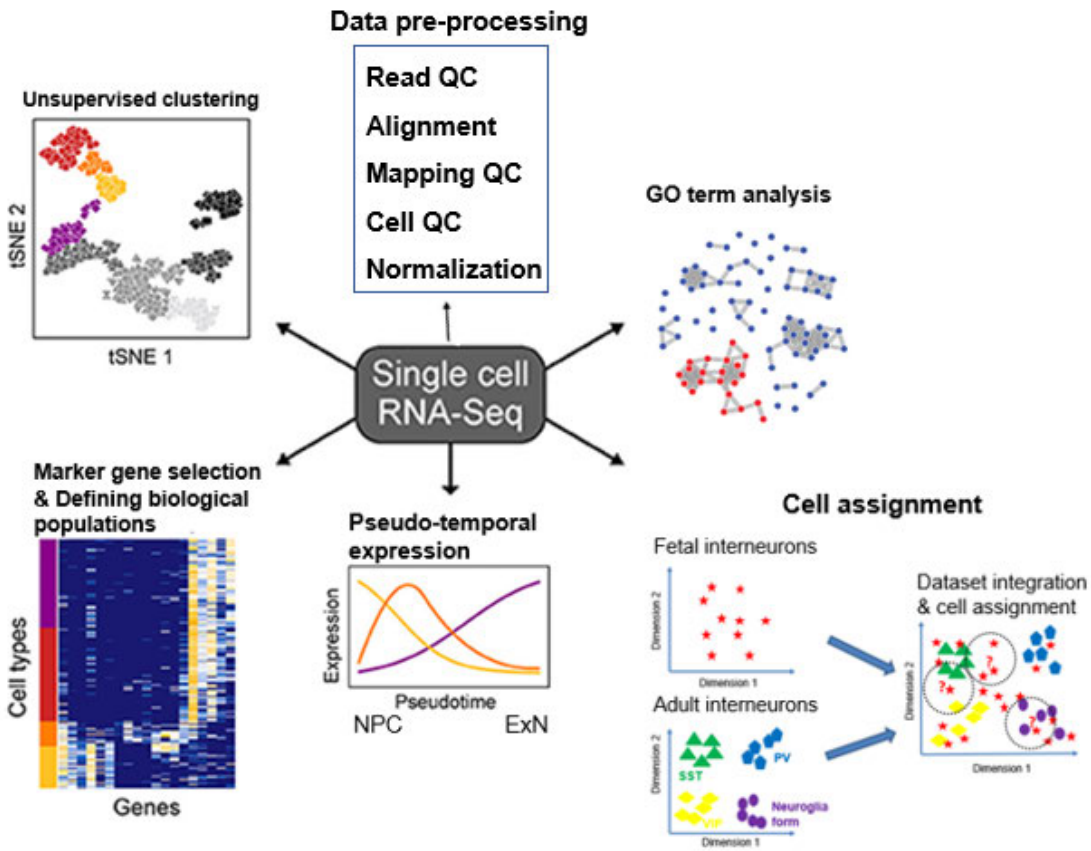


Figure 10: Overview of scRNA-seq data analysis.

scRNA-seq allows for the study of heterogeneous tissue by measure the transcriptional expression profile of individual cells. Such data can be used to unsupervised cluster cells, classify cell clusters by identifying key marker genes, build differentiation trajectories through development, reveal correlation between embryonic cell types and adult cell types, and elucidate enriched gene ontology in each cell type. Figure is modified from https://www.frontiersin.org/files/Articles/353801/fnins-12-00315-HTML/image_m/fnins-12-00315-g001.jpg.

2.2.1 Data pre-processing

For the published datasets, including Zhong's, Nowakowski's, Lake's, Mayer's, and Tasic's datasets, we downloaded the expression matrix of genes in the original paper, then normalized the data according to how the paper described. For Mi's data, we did the data quality control by ourselves. In addition to the general data pre-processing steps, we implemented a series of quality control measures. First, we counted uniquely mapping reads per cell and used only cells with at least 50,000 unique reads mapped to coding sequences. Next, we checked exonic read distribution, distribution across different chromosomes, GC content distribution and gene expression distribution. Any cell that was 3 standard deviations away from the mean for any of the above mentioned metrics were removed. In total, 366 cells were removed, and 2,658 cells were passed on to downstream analysis. We also filtered gene expression profiles of each cell. Any gene expressed by less than 10 cells at less than 5 counts per million (CPM) was removed. We also removed pseudogenes, miRNA, rRNA, mitochondrial associated and ribosome related genes from further analysis. 13,907 genes were kept for downstream analysis. We used R package *Seurat* to manage our dataset (Satija *et al.*, 2015). Briefly, raw read counts were used to create *Seurat* object followed by log normalization using *NormalizeData* function with scale factor set to 1,000,000. Dataset was then scaled by number of genes expressed.

In order to minimize the effect of cell cycle (CC) in the identification of progenitor cell types in both Zhong's and Mi's datasets, we sought to remove CC from our data through regression. Briefly, we used the default list of CC genes in *Seurat* package and calculated G1/S and G2/M phase scores for each cell using function *CellCycleScoring* from *Seurat*. Then, we calculated the difference between G1/S phase score and G2/M phase score. We performed regression on all cells using the resulted difference.

2.2.2 Dimensionality reduction and unsupervised clustering

We regarded that the highly variable genes are driving heterogeneity across the population of cells. To define highly variable genes (HVGs), we calculated the mean of logged expression values and plotted it against variance to mean expression level ratio (VMR) for each gene. Genes with log transformed gene expression level between 0.5 and 8, and VMR between 0.5 and 5 were considered as highly variable genes.

Then we used principal component analysis (PCA) and t-distributed stochastic neighbor embedding (*t*-SNE) as our main dimension reduction approaches (Van Der Maaten and Weinberger, 2012). PCA was performed with *RunPCA* function using HVGs to analyse all cells in each dataset. Following PCA, we conducted jackstraw analysis to identify statistically significant principal components (PCs) that were driving systematic variation. We used *t*-SNE to present data in two-dimensional coordinates. In most analyses significant PCs identified by jackstraw analysis were used as input. Perplexity was set to 30, except when noted otherwise. *t*-SNE plots were generated using R package *ggplot2*. Clustering was done by Luvain-Jaccard algorithm in *Seurat* package using t-SNE vectors.

2.2.3 Differential expression analysis

All differential expression (DE) analyses were conducted using *Seurat* function *FindAllMarkers*. In brief, we took one group of cells and compared it with the rest of the cells, using a Wilcox model. For any given comparison we only considered genes that were expressed by at least 33% of cells in either population. Genes that exhibit p values under 0.05, as well as log fold change over 0.33 were considered significant. All heatmaps plotted using R package *pheatmap*. To define cell clusters in our analysis, we first curated a list of

established marker genes from literatures. Only the genes involved in the list of significantly differential expressed genes were used to perform heatmap or violin plots. Through this process we were able to identify a refined list of genes that were indicative of cellular diversity, specific to our dataset.

2.2.4 Developmental trajectories

The Monocle2 package was used to analyse single cell trajectories in order to discover developmental transitions from NPCs to ExNs in Zhong's dataset (Qiu *et al.*, 2017). We used significantly differentially expressed genes identified by *Seurat* to sort cells in pseudo-time order. The actual gestational time of each cell informed us of the start point of the pseudo-time in the first round of *orderCells*. We then set this state as the *root_state* argument and called *orderCells* again. *DDRTree* function was applied to reduce dimensions and the visualization functions *plot_cell_trajectory* was used to plot the minimum spanning tree on cells. At last, the expression pattern of genes was plotted by *plot_genes_in_pseudotime* function.

2.2.5 Cell assignment between clusters

Since some marker genes of adult interneuron cell types were not expressed in embryonic interneurons, we tried to classify embryonic neurons using information from adult cortical interneurons. To classify embryonic neurons, we utilized two publicly available datasets of adult GABAergic interneurons (Lake *et al.*, 2016; Tasic *et al.*, 2018). We did two times of interneuron assignment between adult and embryonic interneuron cell types. For the interneurons in Zhong's dataset, we compared their transcriptional profiles and cell information with the adult human interneurons in Lake's dataset. For the

interneurons in Mi's dataset, we compared their transcriptional profiles and cell information with the adult mouse interneurons in Tasic's dataset.

We used the same method to deal with the human and mouse datasets. We first found all the shared HVGs in both embryonic and adult datasets. Then we performed feature selection by Random Forest (RF) within the shared HVGs that best represents each cell type for all interneuron cell types defined by the adult human and mouse datasets, referred hereon as the adult cell type features. We then conducted canonical correlation analysis (CCA) on embryonic and adult single cell datasets using the adult cell type features (Butler *et al.*, 2018).

We used the random forest (RF) feature selection and classification technique to get the input genes for the CCA. All the functions described here were retrieved from R package *randomForest*. To conduct RF feature selection, we started with a list of HVGs as an initial set of identifiers for the particular biological process of interest, referred to as features before feature selection (FBFS). FBFS is typically used to define a tentative identity of each sample in question, unless the identity is defined by other metrics. FBFS and tentative identities are then used as the input for R function *randomForest*. We assessed the importance of each FBFS using the importance function and ranked FBFS in descending order. Subsequently, we performed 10-fold cross validation of feature selection on the input genes using the *rfcv* function, with step size set 0.75. The number of features that produces the least error is recorded, and the top n features from the importance measure are regarded as features (FS) and are used in the downstream analysis.

We performed t-SNE analysis to reduce the embryonic and adult cell data onto the same two-dimensional space. Subsequently, we used the two t-SNE vectors for adult cells to conduct k-nearest neighbours analysis (knn) of the adult cell types and reassign cell identities for adult cells (k = 30) using the knn.cv function from R package *FNN*. Briefly, we calculated the average distance ($\bar{d}$) and standard deviation (σ) of all pairs of data points within each 30-cell neighbourhood and removed any neighbour that was more than $\bar{d} + \sigma$

away. Among the remaining neighbours, we counted the identities represented by the neighbours. A cell was assigned the identity represented by the majority, and at least 10, of its neighbours. Through this process, we were able to confidently re-establish the cell types for the adult dataset. The cells that were not able to be assigned were removed from downstream analysis. Subsequently, we used the same *knn* approach on embryonic single cells, using adult cells as neighbours (k = 5) and assigned prospective identities to embryonic cells. We repeated the process again with only the assigned cells as neighbours (k = 5). Each cell was assigned to the most represented identity in its neighbourhood. Through this process we were able to assign embryonic cells to the adult cell types with high confidence.

We then took another independent approach to define the similarity between embryonic neurons and the adult cell type features. *MetaNeighbor* analysis was performed using the R packge *MetaNeighbor* with default settings (Crow *et al.*, 2018). The results from the *MetaNeighbor* analysis were plotted as a heatmap using the R function *heatmap.3*.

2.2.6 Gene Ontology (GO) enrichment analysis

We performed GO enrichment analysis using the *clusterProfiler* package in R (Yu *et al.*, 2012). All parameters were set as default except the pvalueCutoff and qvalueCutoff was set as 0.05. We used differentially expressed genes as the inputs for *enrichGO* function. Significance threshold was set as 0.05 and all other parameters were set as the default settings.

Chapter 3: Vulnerable cell types underlying autism spectrum disorders in the developing human prefrontal cortex

3.1 Introduction

The previous studies by SFARI generated new insight into the causes of ASD and the recent research suggested that ASD may be associated with abnormal brain development during the first few years of growth (Ziats, Edmonson and Rennert, 2015). Based on the bulk mRNA sequencing datasets, the scientists examined the expression levels of SFARI genes across several developmental stages based on bulk mRNA sequencing (Parikshak *et al.*, 2013). They constructed gene expression networks based on the co-expression topological overlap of genes throughout developmental stages, and these networks represented genome-wide functional relationships during fetal and early postnatal brain development. Then they mapped ASD risk genes to these networks. As a result, they found that ASD genes were co-expressed in modules that implicate distinct biological functions during human cortical development, including early transcriptional regulation and synaptic development.

Several molecular or cellular mechanisms lead to the diversity of cortical neurons during the development of human fetal brain, including neurogenesis, differentiation, migration, transcriptional regulation and synaptic function. It was important to consider that the expression pattern of genes in these mechanisms maybe dynamic across the different cell groups, as well as the different developmental stages. For example, based on the co-expression analysis of bulk mRNA sequencing data from the different developmental stages and brain regions, the scientists found that the genes located on *16p11.2* locus were co-expressed with different brain-expressed human

genes. In the comparison of spatiotemporal networks across different brain regions within the same developmental period and across different developmental periods within the same brain region, both the co-expressed interacting partners of genes on *16p11.2* locus and the number of co-expressed pairs between these genes were changed. In detail, *KCTD13* gene, a gene located on *16p11.2* locus, was co-expressed with *CUL3* gene in the inner cortical plate, and the *KCTD13-Cul3* pathway affects the regulation of Rho-GTPase signalling at synapses. Comparison of spatiotemporal networks across different brain regions revealed that the expression levels of *KCTD13* and *CUL3* were positively correlated in the inner cortical plate. Comparison across different developmental periods within the same brain region showed that a higher number of pairs between these genes were found to be co-expressed with *KCTD13* and *CUL3* genes in the cortex during late mid-fetal periods (Lin *et al.*, 2015).

Overall, the scientists found that the expression pattern or co-expression networks of ASD risk genes are dynamic during human brain development by bulk RNA sequencing. However, in general, these studies required more starting material than was available in an individual cell, limiting their application to cell populations. The considerable genetic and cellular heterogeneity among cell types had complicated efforts to establish the biological foundations of ASD. Thus, while such studies had provided important advances, it was becoming clear that the profiling of individual cells would be highly advantageous. The rapid development of scRNA-seq technology added transcriptomic profiling of individual cells to the research area of ASD-related genomic landscape. It can help us reveal new insights into the brain development that establish neuronal identity during development, differentiation, activity, as well as disorders.

3.2 Aim of this chapter

In this chapter, we aimed to identify essential cell types underlying the development of ASD during human brain development by re-analysing published scRNA-seq datasets. The cell types which expressed higher levels of ASD risk genes were regarded as essential cell types of ASD. We mainly focused on two sets of candidate ASD risk genes from the SFARI database, including 86 high confidence ASD risk genes (monogenic genes) and 29 genes at the *16p11.2* locus. We compared the expression pattern of ASD risk genes based on two different steps. Firstly, we compared the expression pattern of ASD risk genes across different cardinal cell classes within the scRNA-seq dataset. Secondly, we performed unsupervised clustering of neuronal cell classes (progenitor cells, excitatory and inhibitory neurons), then compared the expression pattern of ASD risk genes across the cell clusters in each cardinal cell class. After comparison, we evaluated common biological processes converged on cell clusters with enriched expression patterns of ASD risk genes.

3.3 Materials and methods

The published dataset we used in this Chapter was a scRNA-seq dataset which stored in the Gene Expression Omnibus (GEO) under the accession number GSE104276 (Zhong *et al.*, 2018). Zhong et al. used scRNA-seq (Smartseq2, pair-end reads) to identify the molecular signatures that mark cellular diversification located in the prefrontal cortex (PFC). Single cells were collected by mouth pipette. This approach could ensure the cells were always collected individually and preserve the cell viability. In the original paper, the authors classified and identified distinct cardinal cell classes to underlie the development of the human PFC. In this chapter, we used the authors' original classification result of cardinal cell classes. The transcript counts of each cell

were normalized to transcript per million (TPM), where TPM is the transcript count of each gene divided by the sum of transcript counts of that cell, multiplied by one million. In order to minimize the effect of cell cycle (CC) in the identification of progenitor cell types, we removed the effect from progenitor cells by *CellCycleScoring* function in *Seurat* package (Satija *et al.*, 2015). Briefly, we used a published list of CC genes and calculated G1/S and G2/M phase scores for each cell. Then Seurat models the relationship between gene expression and the G1/S and G2M cell cycle scores. The scaled residuals of this model represent a 'corrected' expression matrix, that can be used in downstream analysis.

Due to the potential diversity of cardinal cell classes, unsupervised clustering was conducted in three cardinal cell class, including neural progenitor cells (NPC), excitatory neurons (ExN) and interneurons (IN). The three cardinal cell classes were processed through the same procedure. Firstly, the TPM expression matrix of cells were used to create *Seurat* object followed by log normalization using log (TPM/10+1). Then only protein coding genes that present in at least 0.5% of the cells were used to do clustering. To define highly variable genes (HVGs), we calculated the mean of logged expression values and plotted it against variance to mean expression level ratio (VMR) for each gene. Genes with log transformed mean expression level between 1 and 10, and VMR higher than 0.5 were identified as HVGs. Next, PCA was performed with *RunPCA* function using HVGs to analyse all progenitor cells and statistically significant principal components (PCs) were identified by *Jackstraw* function in *Seurat* (Satija *et al.*, 2015). These significant PCs were used as input to further dimensional reduction. We used t-SNE to present data in two-dimensional coordinates and clustering was done by Luvain-Jaccard algorithm using t-SNE vectors. Finally, all the plots in t-SNE space were generated using *ggplot2* package. Differential expressed genes (DEGs) between clusters were calculated using Wilcox method. For any given comparison we only considered genes that were expressed by at least 33% of cells in either population. Genes that exhibit p-values under 0.05 and log2 fold change values over 0.3 were considered significant.

In order to define biological meaning and developmental states of interesting cell clusters, multiple bioinformatics tools were used for different tasks. For progenitor cells, the *Monocle2* package was used to order the progenitor cells along developmental trajectories and reveal the expression pattern of ASD risk genes during developmental transitions (Qiu *et al.*, 2017). Briefly, we used DEGs identified across progenitor cell clusters as an input gene list to sort cells in pseudo-time order. Then dimensional reduction was performed with *DDRTree* function using DEGs above. Finally, progenitor cells would be mapped to the trajectory which calculated by minimum spanning tree algorithm in the *orderCells* function. The functions called *plot_cell_trajectory* and *plot_genes_in_pseudotime* were used to plot the cells in the dimensional reduced space. For interneuron cells, the biological definition of interneurons was conducted by canonical correlation analysis (CCA) and *MetaNeighbor* package. CCA is a multivariate model based on linear associations between two sets of variables for finding maximum correlation (Butler *et al.*, 2018). Here the CCA algorithm was used to analyse statistical correlations between embryonic interneuron cells in this dataset and adult interneuron cell types in Tasic's dataset which is listed in Table 2. Then the gene expression matrix was scaled based on the correlations to avoid batch effects or bias in normalization procedures between two datasets. After scaling, we performed t-SNE plot to show the embryonic and adult interneuron cells in the same two-dimensional spaces. *MetaNeighbor* was used to reveal the sets of variably expressed genes which can identify cell clusters with high accuracy across embryonic interneuron clusters (Crow *et al.*, 2018). The GO term enrichment analysis of a gene set was performed by *clusterProfiler* package (Yu *et al.*, 2012). Go terms that exhibit adjust p-values under 0.05 were considered significant. All parameters in packages and algorithms were set as the default settings except the parameters we introduced elsewhere.

3.4 Results

3.4.1 Cellular heterogeneity in the developing human prefrontal cortex

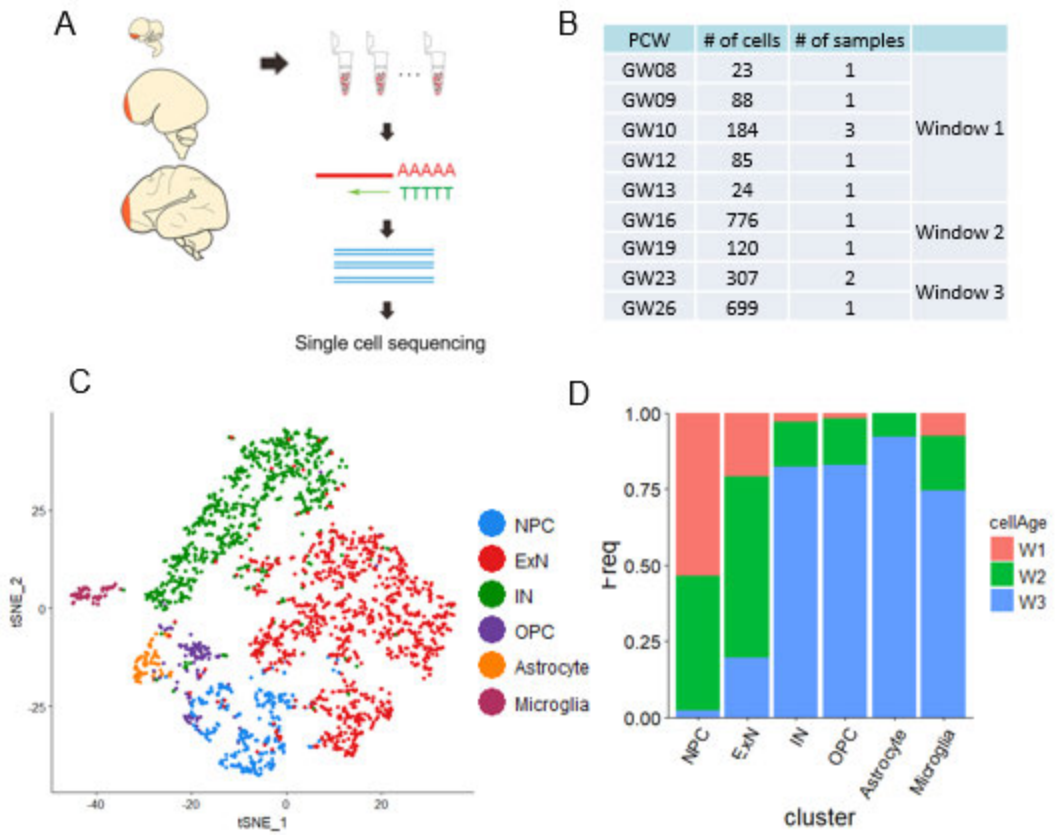
As described in Chapter 1, over the course of cortical development, a number of distinct cell classes are generated. The expression levels of marker genes across different cell classes are variable, and the expression levels in the same cell type are also not static throughout their lifetime (Ohtaka-Maruyama and Okado, 2015). In order to identify the expression patterns of ASD risk genes among the different cell classes across developmental trajectory where disease risk might converge, we applied bioinformatics analysis of scRNA-seq data to define cardinal cell classes of the human prefrontal cortex and calculate the differentially expressed genes across these cardinal cell classes.

Based on a recently published scRNA-seq dataset, we analysed the expression levels of protein coding genes among 2,307 cells (Figure 11A). These prefrontal cortical cells were collected from early to mid-gestation (GW8 to GW26) (Figure 11B). Over this period the major germinal zones and the developing cortical laminae contain migrating and newly born neurons, and neurodevelopmental processes occurring during this period are implicated in neurodevelopmental disorders. As described in Chapter 1, the cortical development stages in this dataset was divided into three windows based on milestones including neurogenesis, differentiation, migration and synaptic function (Kang *et al.*, 2011; Willsey *et al.*, 2013).

For the expression levels of genes, TPMs were obtained from the original paper. Six cardinal cell classes were revealed in this dataset as the authors' original classification of cell classes: neural progenitor cells (NPCs), excitatory neurons (ExNs), interneurons, oligodendrocyte progenitor cells (OPCs), astrocytes, and microglia. In order to plot all cells, PCA was performed by *RunPCA* function using DEGs across six classes. Following PCA, statistically significant principal components (PCs) that were driving systematic variation

were identified. Then significant PCs identified by jackstraw analysis were used as input to draw two-dimensional coordinates of tSNE space. Finally, visualization of the cardinal cell classes was coloured in t-SNE space and dots indicated individual cells (Figure 11C).

To comprehensively study the contribution of cell class composition changes in the human brain temporal transcriptome, we dissected the divergent proportions of developmental windows (Ws) across cell classes. We identified multiple groups of cells at different stages of neuronal differentiation and maturation, corresponding to all known cardinal cell classes at this developmental period. Histograms illustrate the relative contribution of Ws to each cardinal cell class in this dataset (Figure 11D). Most of the cells that captured from W1 were identified as NPCs. ExNs were consisted by a majority of W2 cells and a part of W2 and W3 cells. Cortical interneurons in this dataset were captured from W3. Most of glia cells, such as OPC, Astrocyte and Microglia, were captured from W3 as well. For the neuronal cell classes, the distribution of developmental windows fitted our expectation as we discussed in Chapter 1. At W1, only NPCs were existed, then early deep layer (DL) and late upper layer (UL) neurons were sequential generated. The early born neurons in the GE are referred to as interneurons and travel tangentially within the marginal and intermediate zones during W2 and W3. We do not know the detailed time points about the differentiation and migration of glia cells.



PCW	# of cells	# of samples	
GW08	23	1	Window 1
GW09	88	1	
GW10	184	3	
GW12	85	1	
GW13	24	1	
GW16	776	1	Window 2
GW19	120	1	
GW23	307	2	Window 3
GW26	699	1	

Figure 11: Overview of developing human prefrontal cortex.

(A) Experimental workflow for scRNA-seq of human developing prefrontal cortex (Zhong et al. 2018). (B) Table summarizing developmental windows of brain samples. (C) *t*-SNE plot showing the cardinal cell classes in the dataset. (D) Bar plot depicting the percentage of developmental windows in each cardinal cell classes.

3.4.1.1 Expression pattern of ASD risk gene in cardinal cell classes of human fetal cortex

These cell clusters showed distinct cardinal class aggregation and specific gene expression profiles associated with neuronal classification (Figure 12A). A list of well-known cell class markers was used to illustrate the classification across six cardinal cell classes (Camp *et al.*, 2015; Pollen *et al.*, 2015; Nowakowski *et al.*, 2017). *PAX6*, *HES1* and *VIM* were used as markers to identify NPCs. *NEUROD2*, *NEUROD6* and *RBFOX1* were markers of ExNs. *GAD1*, *GAD2*, *DLX1* and *DLX2* were widely used markers of INs. *OLIG1*, *OLIG2* and *COL20A1* were OPC markers. *GFAP*, *AQP4* and *SLCO1C1* were used as markers to identify the astrocyte. *PTPRC* and *P2RY12* were used as markers of microglia. The expression pattern of these marker genes shown that the cells were correctly identified.

Here we revealed the expression pattern of monogenic ASD risk genes and CNV genes on *16p11.2* locus across six cardinal cell classes in the developing human brain (Figure 12B, 13 and 14). Seventeen monogenic genes as well as seven CNV genes on *16p11.2* locus were included in the DEGs across six cardinal cell classes (Figure 12B). From the heatmap, we observed most of monogenic genes were enriched in INs, and three out of the seven CNV genes were enriched in NPCs. The IN was regarded as an essential cell class for ASD since most of differentially expressed ASD risk genes enriched in this class. The three genes on *16p11.2* locus, *PPP4C*, *HIRIP3* and *KIF22*, were enriched in NPC. The roles of these genes played in NPC will be discussed later in this Chapter.

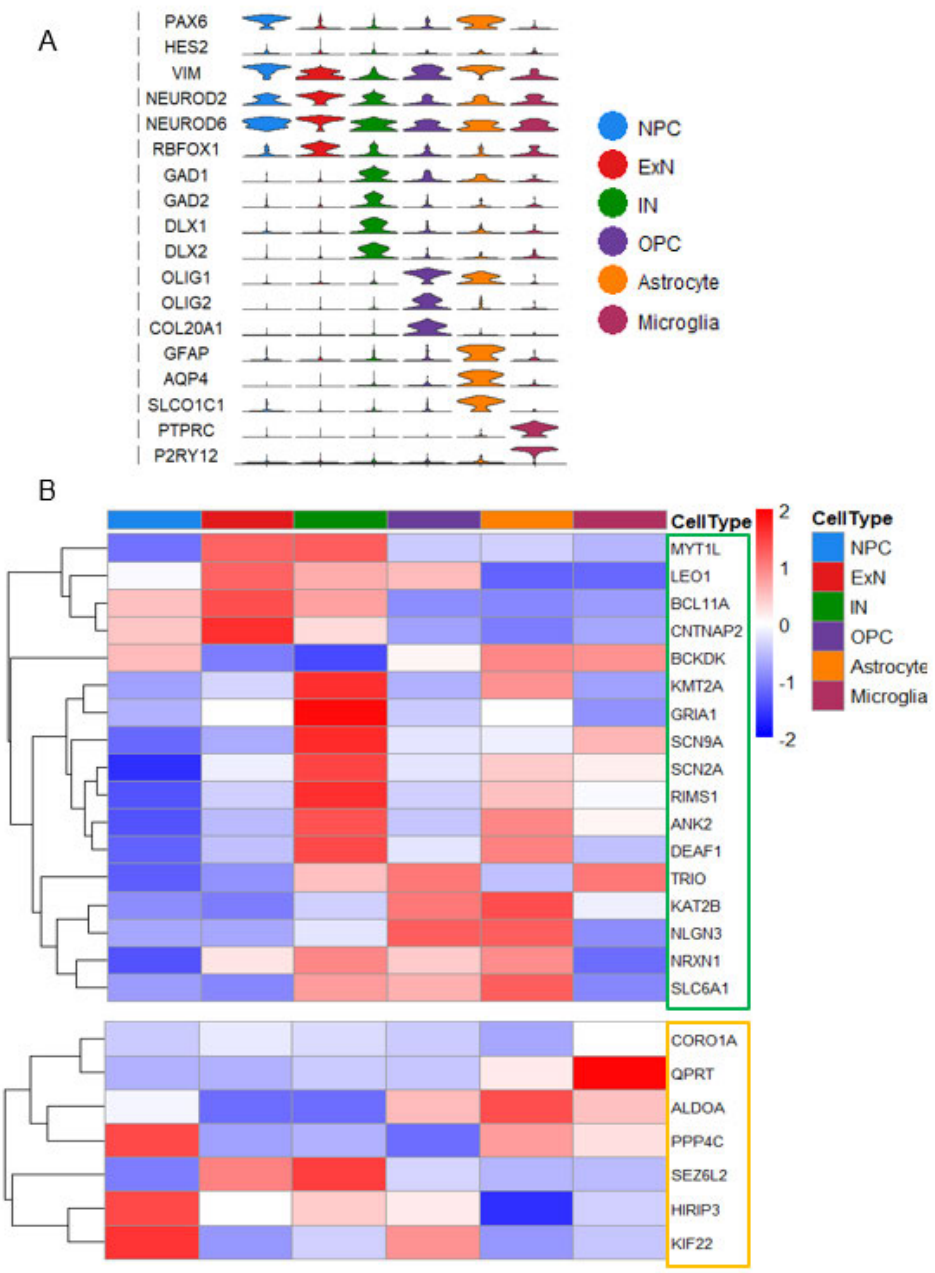


Figure 12: Transcriptional heterogeneity among cardinal cell classes from human fetal cortex.

(A) Violin plot illustrating the expression pattern of marker genes of six cardinal cell classes. (B) Heatmap illustrating the expression pattern of significant differentially expressed ASD risk genes across cardinal cell classes. Green box: monogenic ASD risk genes; yellow box: CNV genes on *16p11.2* locus. The mean expression of genes are centred and scaled from -2 to 2.

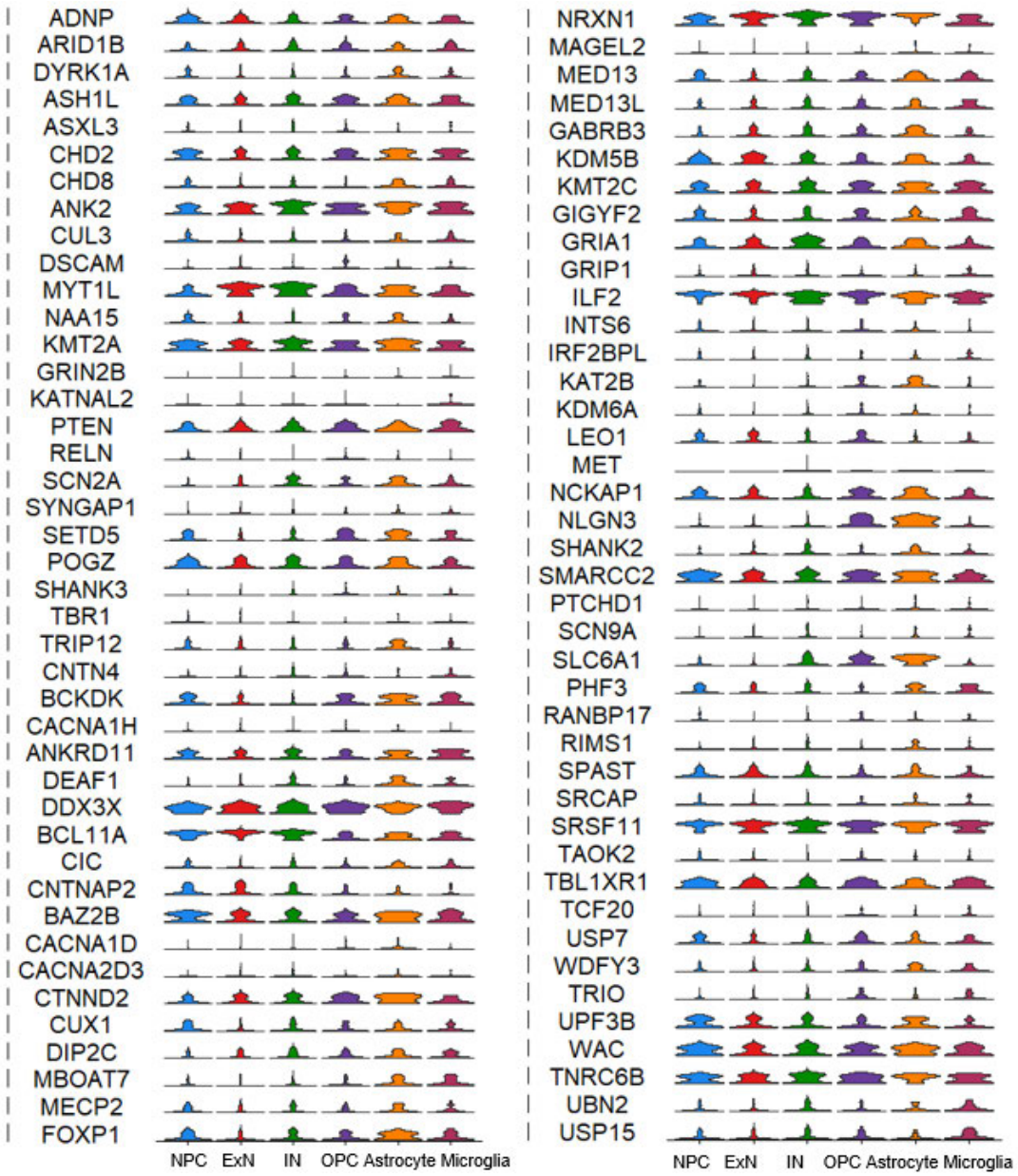


Figure 13: Violin plot illustrating the expression pattern of monogenic ASD risk genes among six cardinal cell classes.

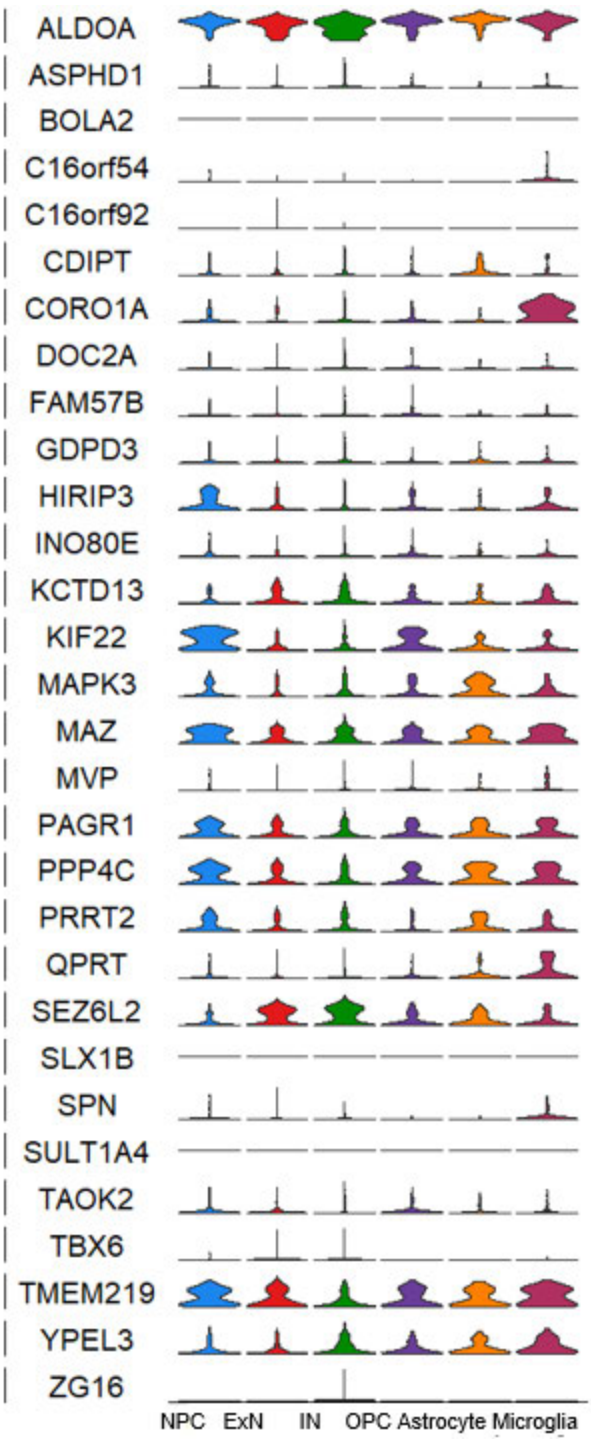


Figure 14: Violin plot illustrating the expression pattern of CNV genes on *16p11.2* locus among six cardinal cell classes.

3.4.1.2 The variances of ASD risk genes expression in each cardinal cell class

An advantage of the single cell approach is that the gene expression levels can be investigated not only on mean expression values, but also across the cell population. By studying the distribution of gene expression levels across the population, the detailed cell-to-cell variability in gene expression can be revealed (Grün *et al.*, 2015). As we described in Chapter 1, in each cardinal cell classes, there are several distinct cell subpopulations. We plotted the expression levels of three ASD risk genes to illustrate if the expression levels of these genes were consistent within each cardinal cell class (Figure 15). As a result, we found that *KIF22* gene was highly expressed in most of NPCs, but *FEZF2* gene was highly expressed in the half of ExNs, and *SCN9A* genes was only highly expressed in a small part of INs. This result indicated that some ASD risk genes could only effect on a small group of cells in a cardinal cell class. To further reveal the expression pattern of ASD risk genes across the potential subpopulations within each cardinal cell class, we performed a more detailed analysis within each cardinal cell class.

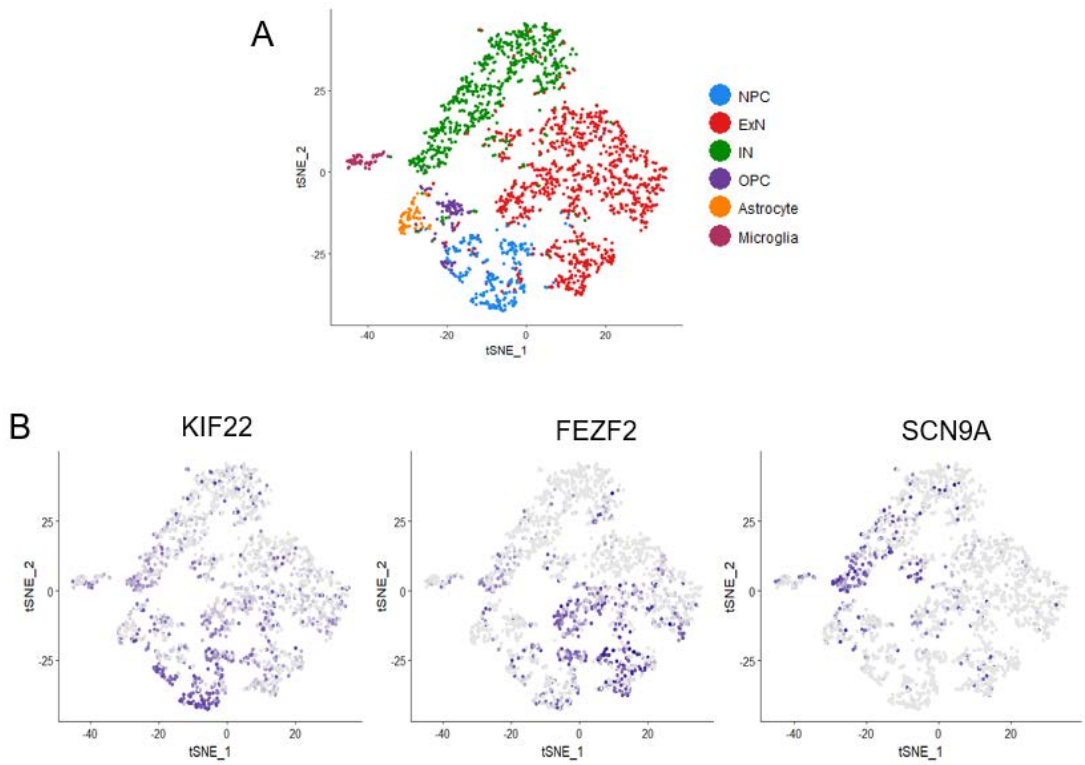


Figure 15: Gradient plots of the expression levels of ASD risk genes in the t-SNE space.

(A) t-SNE plot showing the cardinal cell classes in the dataset. (B) Colour intensity indicating the relative number and level of ASD risk genes expression among the cardinal cell classes. Grey, low expression levels; Blue, high expression levels.

3.4.2 Analysis of cell types within each cardinal cell class

To classify the cell types of neuronal cells in the developing PFC, we performed clustering analysis using Seurat as described in Chapter 2. In detail, PCA and tSNE were used as main dimension reduction approaches. PCA was performed with *RunPCA* function using HVGs to analyse all cells in Zhong's dataset. Following PCA, statistically significant principal components (PCs) that were driving systematic variation were identified. Then significant PCs identified by jackstraw analysis were used as input to draw two-dimensional coordinates of tSNE space. Finally, unsupervised Clustering was done by Luvain-Jaccard algorithm based on *t*-SNE coordinates. As a result, six different cell types of NPC, four different types of ExNs and eight different types of INs were revealed. OPC, Astrocyte and microglia were not further subdivided (Figure 16). The detailed expression pattern of ASD risk genes across these cell types in each cardinal cell class will be discussed next in this Chapter.

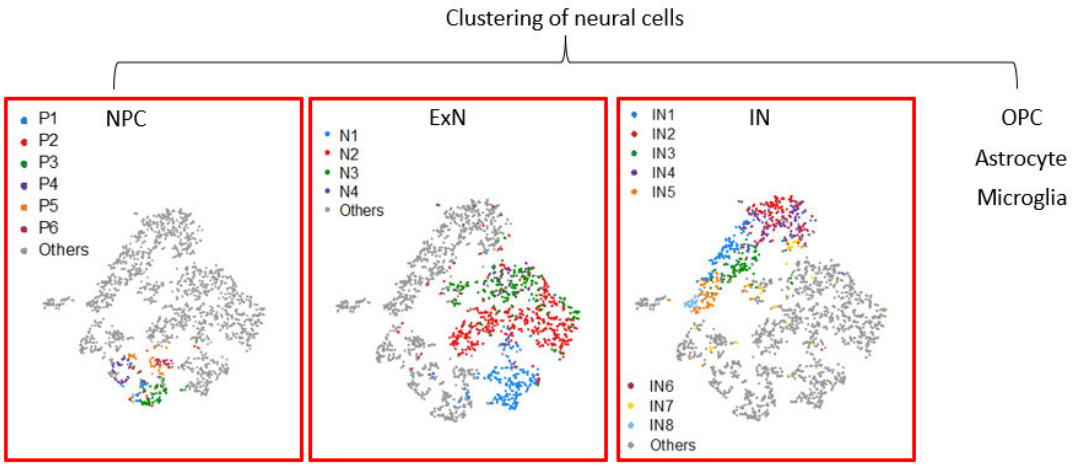


Figure 16: Unsupervised clustering identifies distinct cell types in cardinal cell classes.

Dots indicates individual cells; colour means cell types.

3.4.3 Diversity of cortical progenitors in the human fetal cortex

As described in Chapter 1, NPCs are differentiating into different cell types defined by molecular characteristics and functional diversity. At the beginning, the self-renewal stem cells in the VZ can develop into vRG cells. These early vRG cells can differentiate to pia-contacting oRG that delaminate from the VZ and translocate to the OSVZ. IPCs that located in OSVZ originated from vRG and oRG. All these NPCs (vRG, oRG and IPC) are mitosis cells and featured by distinct marker genes and functions.

We did unsupervised clustering to identify cell clusters of NPCs and calculated the DEGs among clusters, to identify genes that are most informative for defining cell types. Finally, the expression levels of ASD risk genes within each cell cluster were compared.

Six progenitor clusters were identified based on their transcriptional profiling and labelled as P1, P2, P3, P4, P5 and P6 (Figure 16, NPC). Histograms illustrate the relative contribution of developmental windows to each progenitor cell cluster (Figure 17A). The majority of progenitor cells in P1, P2 and P6 were captured from W1. P4 and P5 were mainly consisted by W2 cells with a few W1 cells. Progenitors cells in P3 were equally captured from W1 and W2.

We used a set of well-known maker genes of progenitor cell types to define cell identities of the clusters (Figure 17B). All cells in six clusters were marked by expression of progenitor markers *PAX6*, *VIM* and *HES1*. *HOPX*, *TNC* and *MOXD1*, which have been identified as markers of oRG were high expressed in P4, while *EOMES*, *PPP1R17*, *NHLH1* and *RBFOX1* was expressed in P6, which suggests that the cells in the cluster were IPCs. Notably, both progenitor markers *PAX6* and *VIM*, and some maker genes of IPCs (e.g., *EOMES* and *PPP1R17*) were enriched in P3 cells, whereas *NHLH1* and *RBFOX1* were not. This suggests that P3 may contain both IPCs and vRG cells. P5 represent a mixture of cell types based on marker gene expression, as both markers of oRG (e.g., *HES1, HOPX, TNC* and *MOXD1*) and makers of IPCs (e.g., *EOMES, PPP1R17* and *RBFOX1*) were enriched.

We have identified a set of novel marker genes for these progenitor cluster by differential expression analysis (Figure 17C). We examined the expression pattern of the monogenic ASD risk genes and *16p11.2* genes among six progenitor clusters in the developing human brain (Figure 18 and 19). Five monogenic genes as well as one *16p11.2* gene were included in the DEGs across six clusters (Figure 17D). From the heatmap, we observed most of monogenic genes and the one *16p11.2* gene were enriched in P6, and two out of the five monogenic genes were enriched in P5.

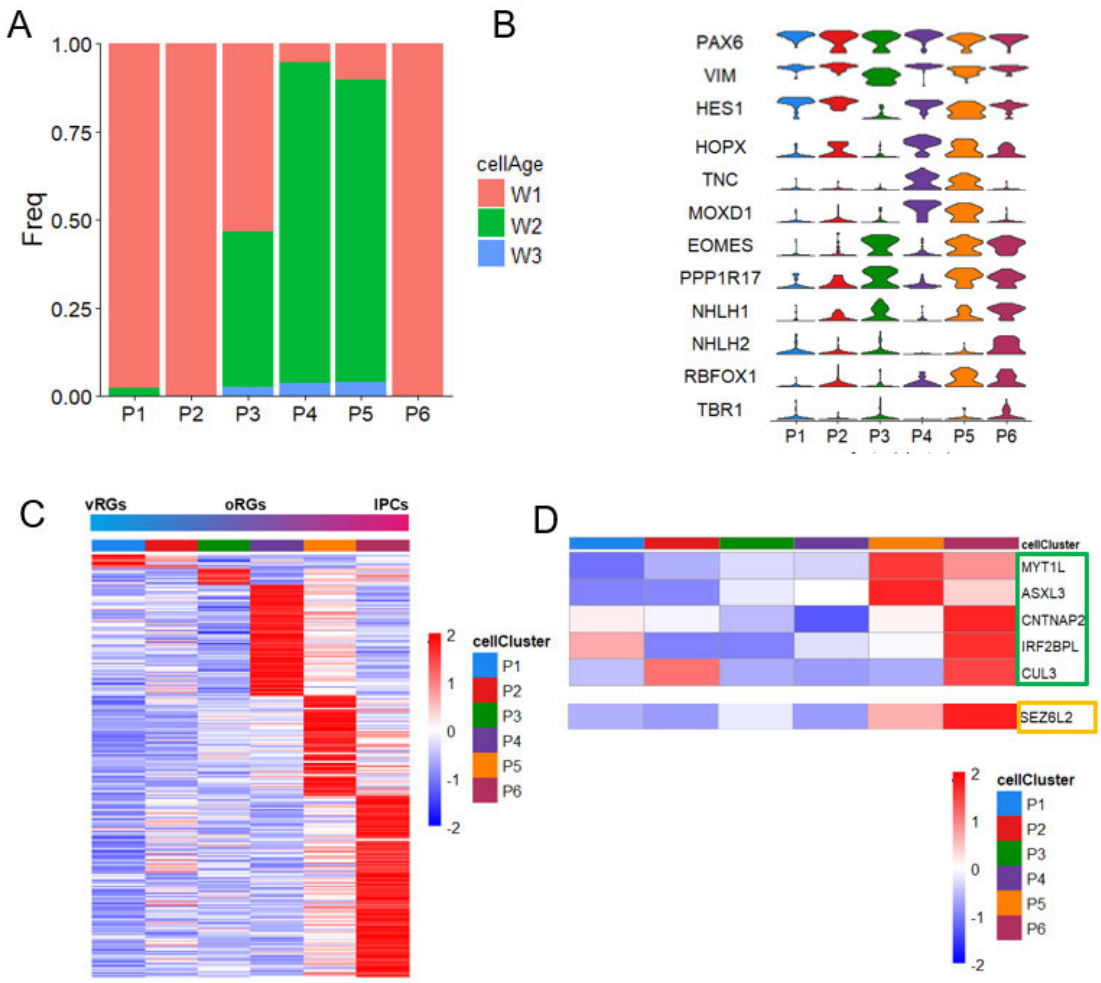


Figure 17: Diversity of cortical progenitor cell types in the human fetal cortex.

(A) Bar plot depicting the percentage of developmental windows in each cell type. (B) Violin plot illustrating the expression pattern of marker genes of cell types in NPC. (C) Heatmap illustrating the expression pattern of differentially expressed genes across cell clusters in NPCs. The ventral radial glia cells (vRGs), outer radial glia cells (oRGs) and intermediate progenitor cells (IPCs) are defined by the expression of known markers as listed in A. (D) Heatmap illustrating the expression pattern of ASD-DEGs across cell clusters. Green box: monogenic ASD risk genes; yellow box: CNV genes on *16p11.2* locus.

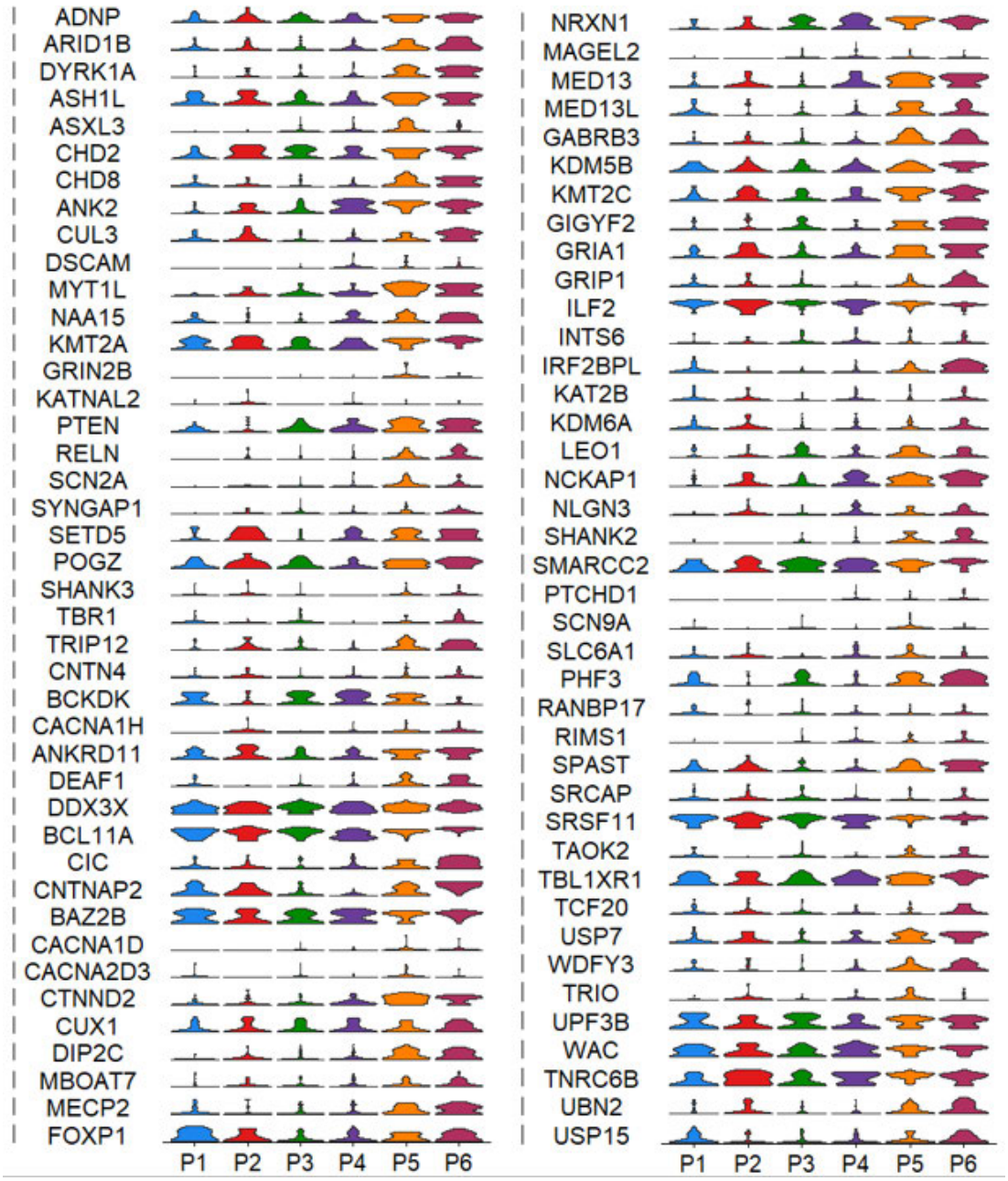


Figure 18: Violin plot illustrating the expression pattern of monogenic ASD risk genes among six NPCs clusters.

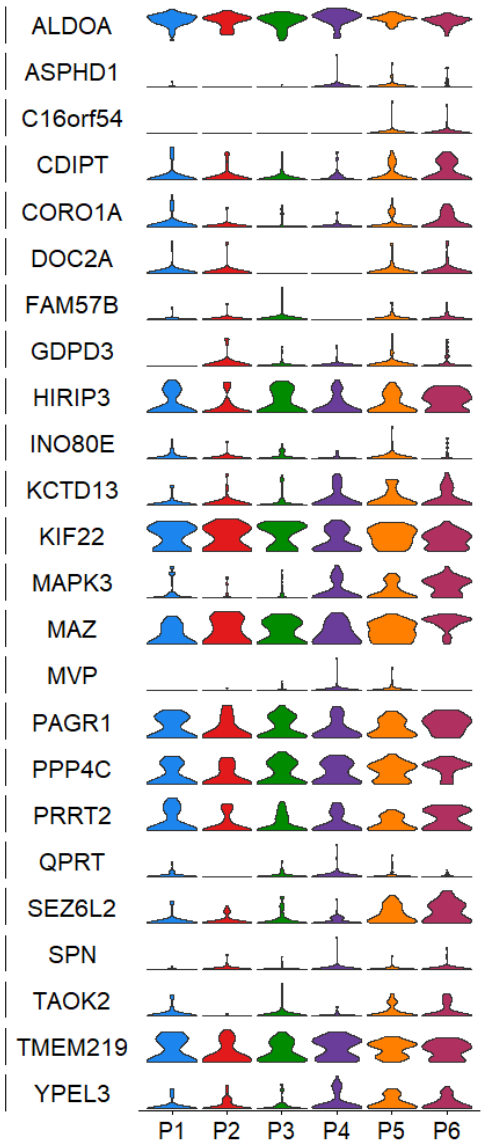


Figure 19: Violin plot illustrating the expression pattern of CNV genes on *16p11.2* locus among six NPCs clusters.

3.4.3.1 Elevated expression of ASD risk genes in intermediate progenitors

The expression of ASD risk genes were enriched in P5 and P6, and IPC marker genes were highly expressed in both P5 and P6. So we wanted to establish what these cells are. It was possible to determine the relative position of each cell over the trajectory, and we can plot the expression pattern of ASD risk genes over the developmental timeline.

The *Monocle2* package was used to infer developmental trajactory among the progenitor cell types, and the expression levels of ASD risk genes were plotted along the developmental trajactory. This analysis revealed a progenitor developmental trajectory that linked vRG cells, through oRG cells to IPCs (Figure 20A). Notably, expression of genes known to be enriched in vRG cells (e.g., *HES1* and *VIM*), oRG cells (e.g., *HOPX* and *MOD1*), and IPCs (e.g., *RBFOX1* and *TNC*) exhibited restricted expression along the pseudo-time (Figure 20).

This trajectory also revealed that the vRG–oRG–IPC lineage correlates with the developmental windows of the developing PFC (Figure 20). vRG cells that collected at W1 were act as the root state of development, oRG cells that collected at W2 were followed by vRG cells over the trajactory, and IPCs which come from W2 may come from both vRG and oRG cells. These results suggest that IPCs might originate from both the vRG cells at W1 and oRG cells at W2.

To illustrate the expression pattern of ASD risk genes, we assessed the expression pattern of ASD-DEGs along the trajectory. The relative expression of five monogenic ASD risk genes as well as one *16p11.2* CNV gene that included in the ASD-DEGs above were enriched at the end of trajectory, where most cells were identified as IPCs at both W1 and W2. Analysis of GO term enrichment in the enriched genes of P6 suggested that these genes participated in multiple processes of neuronal development, such as axon, dendrite and synapse. Terms related with “synapse” were known pathways

about neural development, which means the cells in P6 may be in a trasition state between IPCs and early neurons.

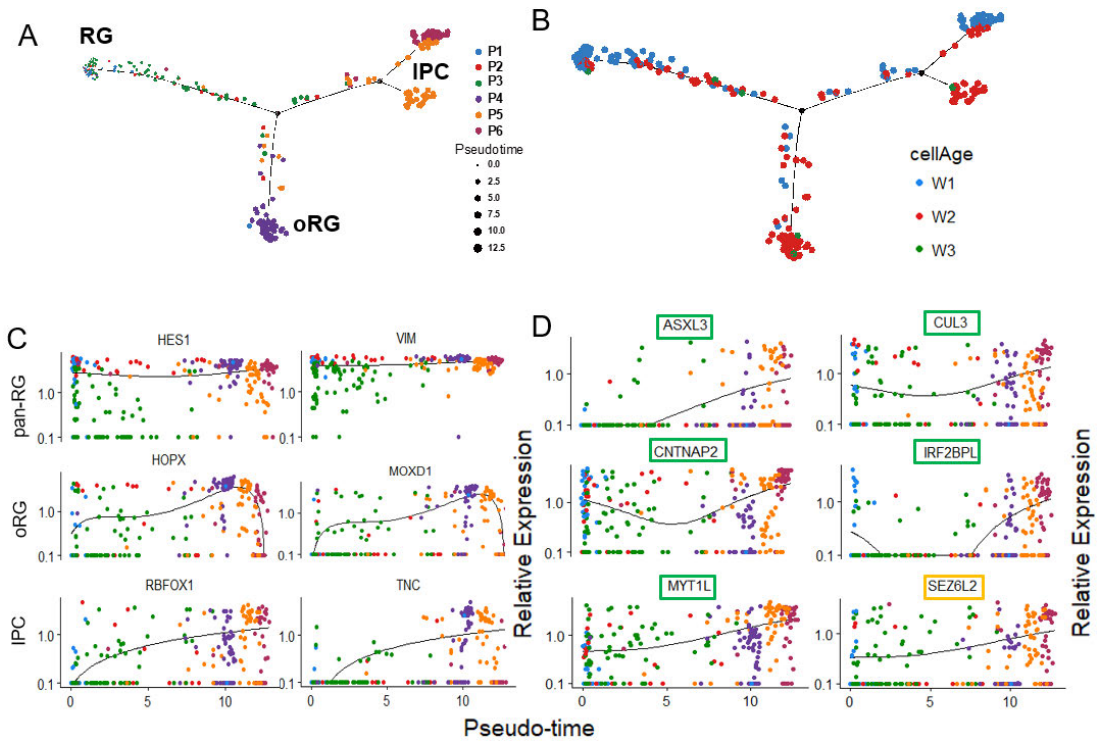


Figure 20: Developmental trajectories of cortical progenitor cells by Monocle2.

(A) *Monocle2* recovered a branched single-cell trajectory. Cells were coloured based on cell type and plotted in the 2D independent component space. (B) Relationship of developmental windows in the NPCs. (C) Expression of known markers with pseudo-time. The colour of cells indicates the cell groups in (A). (D) Expression of ASD-DEGs with pseudo-time. The colour of cells indicates the cell groups in (A). Green box: monogenic ASD risk genes; yellow box: CNV genes on *16p11.2* locus.

3.4.4 Specification of excitatory neurons in the developing human cortex

As described in Chapter 1, during human cortical development, progenitors are generated in the VZ and then migrate radially through the SVZ-IZ in waves to the expanding CP. Excitatory neurons are generated sequentially in an inside-out order from progenitors residing in the VZ and SVZ. This results in the sequential generation of early deep layer (DL) and late upper layer (UL) neurons, as early-born neurons settle in deep layers of the cortex, whereas late-born neurons populate the upper layers. In detail, the adult cerebral cortex is organized into six layers (L2-L6), and the excitatory neurons can be found in all cortical layers except layer I. Previous studies indicated that the excitatory neurons in different layers are expressing distinct marker genes and playing different biological roles. For example, deep cortical layers (L5 and L6) contain neurons that highly express *BCL11B* (also known as *CTIP2*), *TBR1* and *FEZF2* genes. These neurons are called corticothalamic projection neurons and subcortical projection neurons, which project and carry information from the cerebral cortex to subcortical structures including the thalamus (Chen *et al.*, 2008; Greig *et al.*, 2013). The upper layers (L2-L4), also known as superficial layers, contain neurons that mainly express *SATB2*, *RORB* and *CUX1* genes. These neurons are called callosal projection neurons, which project and carry information to contralateral brain regions thereby transmitting information from one brain hemisphere to the other (Kwan, Šestan and Anton, 2012).

In order to predict the laminar location of embryonic ExNs in this dataset, we did unsupervised clustering to identify cell clusters of ExNs and identified the cells in each cluster to either DL or UP based on the expression levels of DL and UL marker genes. Finally, the expression levels of ASD risk genes within each cell cluster were compared.

3.4.4.1 Excitatory neurons classified by unsupervised clustering

Four excitatory neuron clusters were identified based on their transcriptional profiling and labelled as N1, N2, N3 and N4 (Figure 16, ExN). Histogram illustrates the relative contribution of Ws to each excitatory neuron cluster (Figure 21). The majority of excitatory neurons in N1 were captured from W1. N2 were mainly consisted of W2 cells with a few W1 cells. Excitatory neurons in N3 and N4 were mostly captured from W3 with a few W2 cells. Since early-born neurons settle in deep layers of CP, and late-born neurons populate the upper layers, this distribution of Ws indicated most of neurons in N1 and N2 were likely to be DL-like excitatory neurons, and neurons in N3 and N4 likely to be UL-like excitatory neurons. Some maker genes were used to check the layer-specificity of these clusters (Figure 21B). Some deep layer markers, such as *TLE4, SOX5, SSTR2, FEZF2,* and *BCL11B* were high expressed in N1 and N2, while the upper layer markers, such as *CUX2* and *SATB2*, were enriched in N3 and N4. Other upper layer markers, such as *CUX1, UNC5D, RORB, WFS1* and *RELN*, were not enriched in N3 and N4. This means the deep layer markers among embryonic neurons can well define the layer-specificity of DL-like neurons (N1 and N2), but marker genes of upper layer showed limited correspondence to distinguish the maturing neuron clusters (N3 and N4) or UL-like neurons.

A set of novel marker genes were identified for these excitatory neuron clusters by differential expression analysis (Figure 21C). We examined the expression pattern of the monogenic ASD risk genes and genes on *16p11.2* locus among four excitatory neuron clusters in the developing human brain (Figure 22 and 23). Seven monogenic genes were included in the DEGs across four clusters (Figure 21D). From the heatmap, we observed that most of ASD monogenic genes, such as *ASXL3, RELN, BCL11A, CNTNAP2, CTNND2* and *KMT2C*, were enriched in N1 and N2. Only *SCN2A* was enriched in N3. There was no CNV genes differentially expressed among the four clusters. We applied the analysis of GO term enrichment for the enriched genes between DL-like

neurons (N1 and N2) and UL-like neurons (N3 and N4), but there was no significant GO term identified.

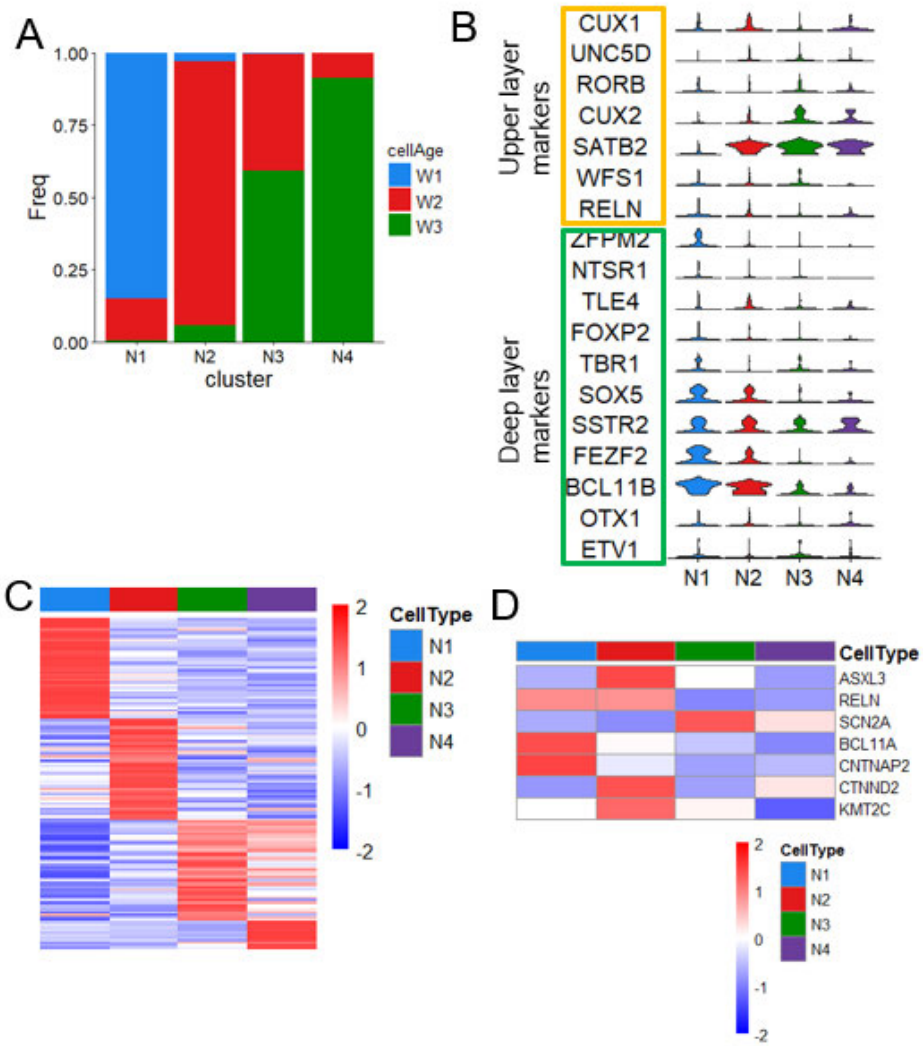


Figure 21: Unsupervised clustering of excitatory neurons in the human fetal cortex.

(A) Bar plot depicting the percentage of developmental windows in each cluster. (B) Violin plot illustrating the expression pattern of marker genes between deep layer and upper layer. Yellow box: upper layer maker genes; Green box: deep layer maker genes. (C) Heatmap illustrating the expression pattern of differentially expressed genes across cell clusters within excitatory neurons. (D) Heatmap illustrating the expression pattern of differentially expressed ASD risk genes across cell clusters.

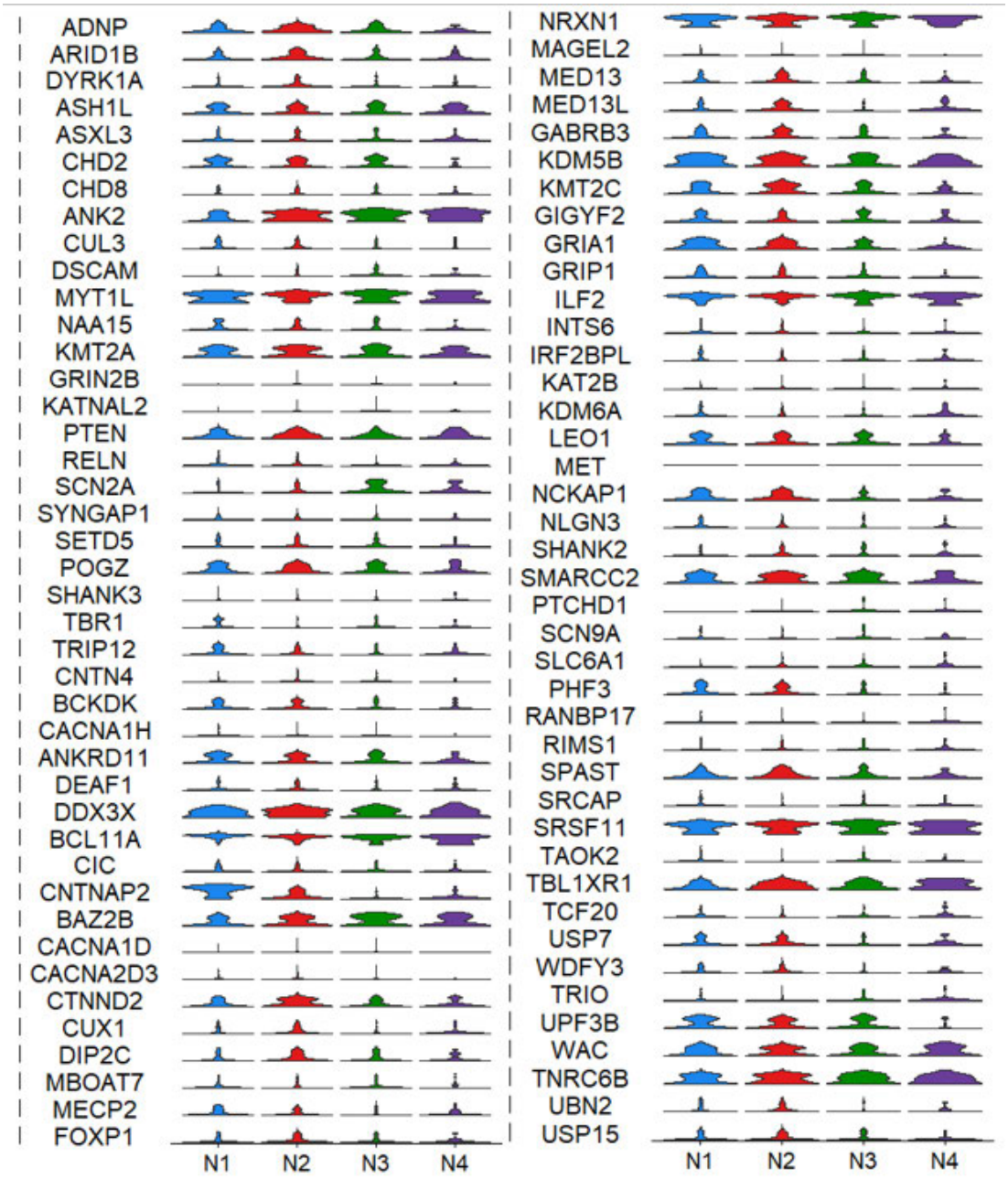


Figure 22: Violin plot illustrating the expression pattern of monogenic ASD risk genes among four ExN clusters.

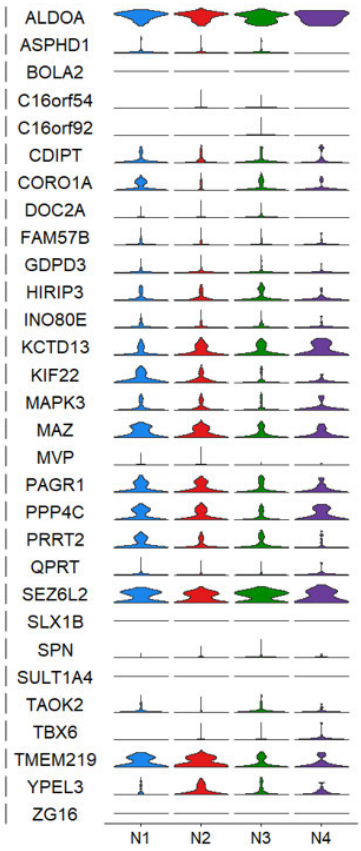


Figure 23: Violin plot illustrating the expression pattern of CNV genes on *16p11.2* locus among four ExN clusters.

3.4.5 Expression pattern of ASD risk genes in interneurons of human fetal cortex

As described in Chapter 1, there was a study that claimed interneurons were primarily affected by genetic susceptibility of ASD (Skene and Grant, 2016a). In this study, we also noticed that many ASD risk genes were enriched in INs, and the expression levels of ASD risk genes were not consistent within INs. In order to explore the potential expression pattern of ASD risk genes within INs, we used unsupervised clustering to identify cell clusters of cortical interneurons and calculated the DEGs among clusters. Then the expression levels of ASD risk genes within each cell cluster were compared to identify clusters of interneurons that expressed ASD risk genes and might therefore be vulnerable to their mutation. Finally, we tried to reveal essential interneuron cell types underlying ASD in the developing human PFC and detect the distinct molecular programs among these cell types.

3.4.5.1 Diversity of interneurons in human fetal cortex

Eight interneuron clusters were identified based on their transcriptional profiling and labelled as IN1, IN2, IN3, IN4, IN5, IN6, IN7 and IN8 (Figure 16, IN). Histograms illustrate the relative contribution of DWs to each interneuron cluster (Figure 24). The majority of interneurons in IN1, IN2, IN3, IN4, IN5 and IN8 were captured from W3. IN6 and IN7 were mainly consisted by W3 cells with a few W2 cells. The mixture of developmental windows in the clusters indicated that the clustering was not affected by the sampling time.

We have identified novel marker genes for these interneuron clusters by differential expression analysis (Figure 24B). We examined the expression pattern of the monogenic ASD risk genes and *16p11.2* genes among the eight interneuron clusters (Figure 25 and Figure 26). Thirty-five monogenic genes

were included in the DEGs among eight clusters and none of *16p11.2* genes were found within the list of DEGs (Figure 24C). From the heatmap, we observed most of monogenic genes within the list of DEGs were enriched in IN8, except *EXPH1* and *ETFB*, which were enriched in IN2 and IN4, respectively.

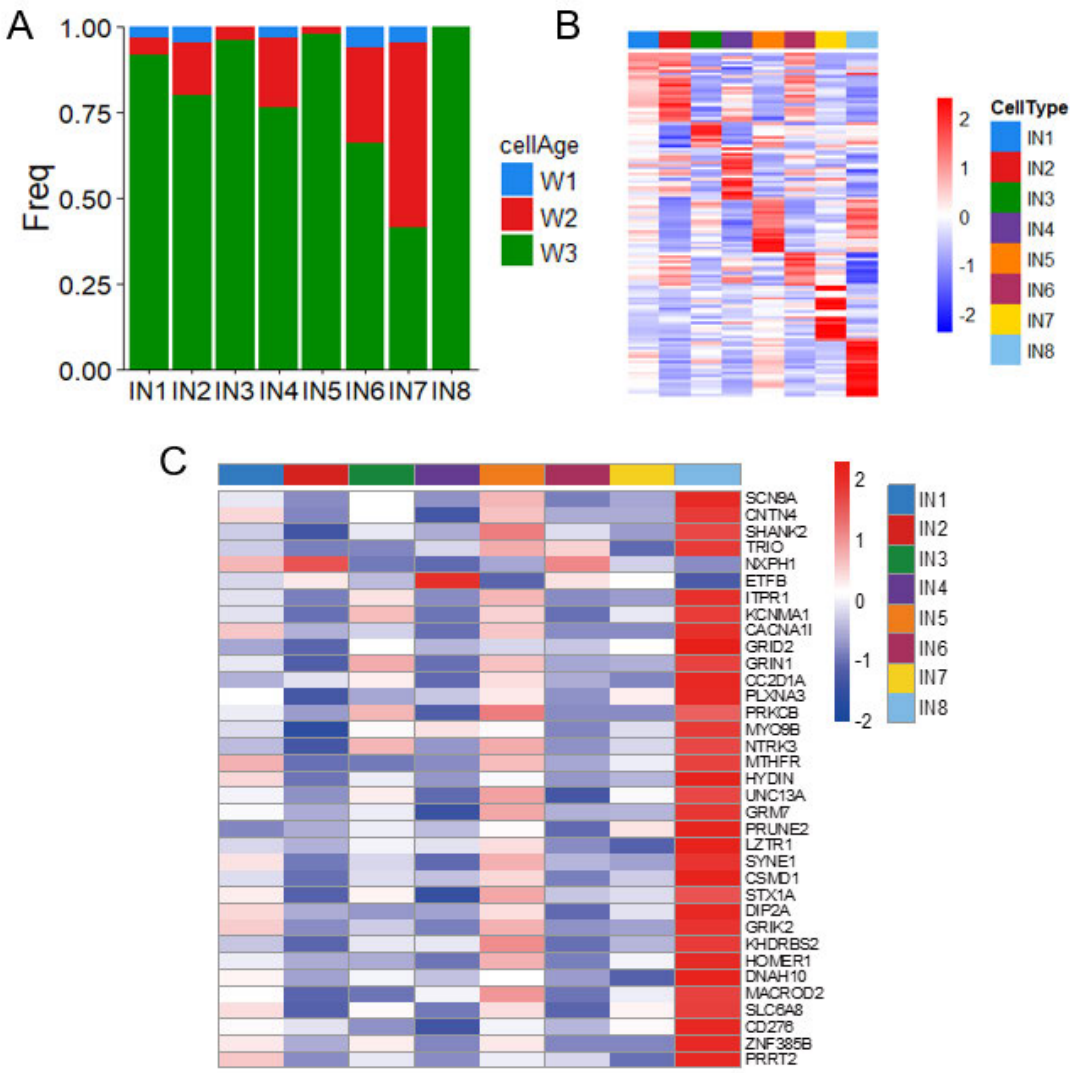


Figure 24: Diversity of interneurons in human developing PFC.

(A) Bar plot depicting the percentage of developmental windows in each cell cluster. (B) Heatmap showing the differential gene expression for each cell cluster in progenitor cells. Red corresponds to high expression level; blue correspond to low expression level. (C) Heatmap illustrating the expression pattern of top differentially expressed ASD risk genes across cell cluster.

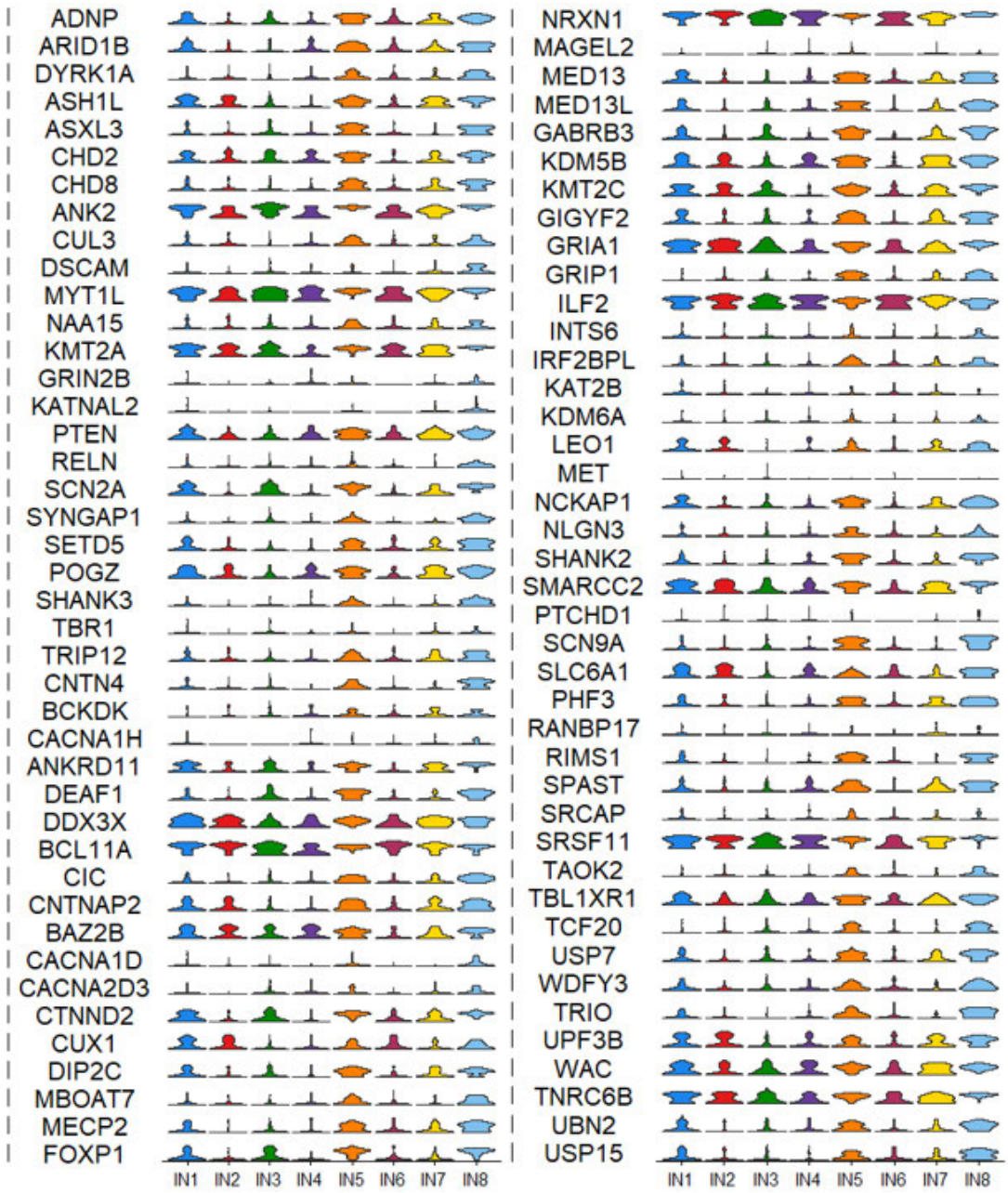


Figure 25: Violin plot illustrating the expression pattern of monogenic ASD risk genes among eight interneuron cell classes.

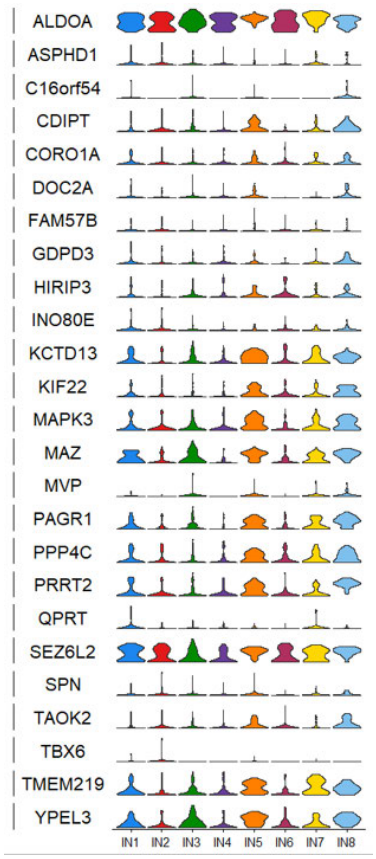


Figure 26: Violin plot illustrating the expression pattern of CNV genes on *16p11.2* locus among eight interneuron cell classes.

Similar to what we did in the analysis of progenitor cells, we used a set of well-known maker genes of interneuron cell types to define cell identities of the cells in IN8. There were sixteen marker genes listed in the violin plot (Figure 27). *LHX6* and *SOX6* genes are marker genes for MGE-derived interneurons. Within the MGE region, *SST, TAC1* and *CALB1* genes regulate the development of MGE-derived SST+, PV+ and CALB1+ cortical interneurons, respectively. *NR2F2, SP8, HTR3A* and *PROX1* genes were marker genes for CGE-derived interneurons. Some marker genes, such as *VIP, CCK* and *ID2* genes, regulate the generation and postnatal maturation of VIP+, CCK+, ID2+ cortical interneurons which migrated from CGE region, respectively. *CALB2, RELN* and *NPY* genes were marker genes of specific interneuron cell types, and these interneurons were migrated from both MGE and CGE regions.

LHX6 and *SOX6* genes, which have been identified as markers of MGE-derived interneurons, were highly expressed in IN1, IN2, IN5, IN6 and IN8. It suggests that most cells in these clusters had come from MGE region. *SST* gene was expressed in cells from IN1, IN2, IN5, IN6, IN7 and IN8, whereas TAC1 gene was low expressing within all clusters. It means MGE-derived interneurons, especially SST+ interneurons, can be found in all clusters except IN3.

The expression of CEG-derived interneuron markers, such as *NR2F2, SP8, HTR3A* and *PROX1* genes, were enriched in IN4. *NR2F2* gene was also high expressed in IN5, IN7 and IN8. *CALB2, RELN* and *NPY* genes were maker genes of some interneuron cell types from both MGE and CGE regions. These genes were highly expressed in multiple clusters, such as IN4, IN7 and IN8. Interestingly, as a CGE-derived interneuron marker, *ID2* gene was high expressed among all clusters. All information above suggests that each interneuron clusters comprise multiple interneuron cell types.

There are two possibilities to explain why mixture of interneuron cell types are found in each interneuron cluster. Firstly, we still know little about the molecular mechanisms regulating MGE and CGE derived cortical interneuron fate. And the understanding of genetic cascade for the interneuron cell fate specification

was not integrated. For example, the well-known marker gene of PV interneuron was *PVALB* gene. However, this gene is not expressed in interneurons at early developmental stages. *TAC1* gene maybe not expressed in interneurons at early developmental stages as well. Secondly, the interneuron clusters were identified based on the unsupervised clustering. The highly variable genes which effect the clustering may not be the molecular markers which specifically label the interneuron cell types. So, the biases cannot be avoided if the interneuron cell types were depicted only based on the expression pattern of these marker genes among clusters.

Analysis of GO term enrichment for the enriched genes in IN8 suggested that these genes participated in multiple processes of neuron-neuron communication, such as postsynaptic regulation, synapse and ion channel (Figure 27). It means that regulation of synapse development could be an important cellular characteristic of IN8 cells.

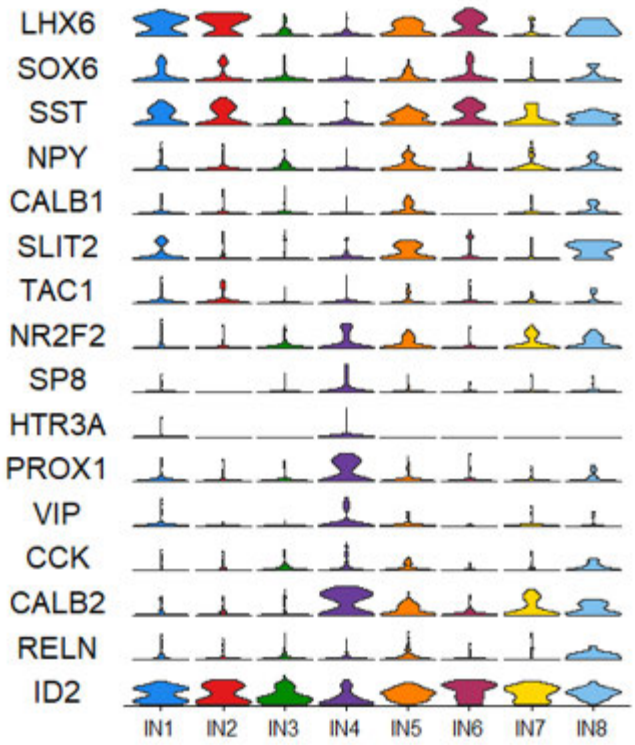


Figure 27: Violin of well-known marker genes of interneuron cell types across clusters.

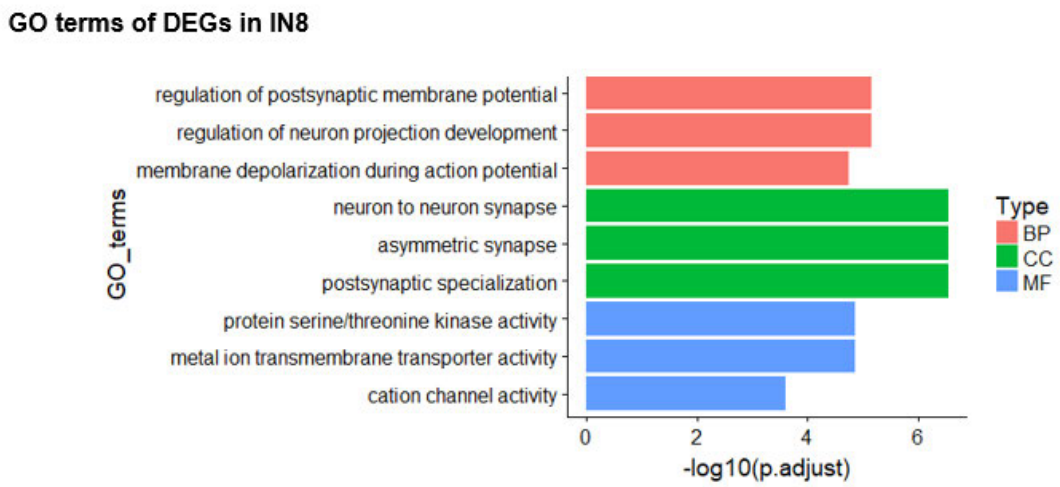


Figure 28: Top significant GO terms associated with the enriched DEGs in IN8.

The vertical axis represents the GO category, and the horizontal axis represents the -log10(adjusted p-value) of the significant GO terms. Greater -log10(adjusted p-value) scores correlated with increased statistical significance. Clusters belong to the different types of terms are color-coded accordingly. BP, biological processes; CC, cellular components; MF, molecular functions.

3.4.5.2 Defining interneuron cell type identity by canonical correlation analysis

Instead of define cell types from unsupervised clustering, the canonical correlation analysis (CCA) in a recent scRNA-seq study has provided new insight to define the interneuron cell type (Butler *et al.*, 2018). By this method, we can combine embryonic and adult datasets together, and infer the cell types of interneurons in embryonic dataset based on the adult cell type features.

A gene expression matrix from the published single-nucleus RNAseq (snRNA-Seq) dataset which contain 10,319 adult interneurons were used as reference dataset in this analysis (Lake *et al.*, 2016). There were four interneuron cell types identified in the adult human PFC dataset: SST, PV, VIP and Neurogliaform. Through computational CCA algorithm, the expression matrixes of human embryonic and adult interneurons were aligned and combined. Firstly, the combined data were represented in a t-SNE space (Figure 29A; top). Then four interneuron cell types defined by Lake and colleagues were labelled in the space (Figure 29A; Middle left). Last, we used the t-SNE coordinates of cells to conduct nearest neighbour's analysis between the embryonic and adult interneurons. An embryonic interneuron was assigned the cell type represented by the majority of the five closest adult interneurons around it (Figure 29; Middle right). Through this process we were able to assign 106 embryonic interneurons to four adult interneuron cell types with high confidence. The marker genes of the four interneuron cell types were plotted to illustrate the alignment result (Figure 29B). Histograms illustrate the relative contribution of interneuron cell types to each cell cluster in the embryonic dataset (Figure 29C). As some embryonic interneurons were far away with the adult interneurons in the t-SNE space, there were lots of interneurons cannot be assigned to any cell type. IN4 were consisted by a majority of VIP+ interneurons and a small part of SST+ and PV+ interneurons. All other clusters, including IN8, were consisted by a majority of SST+ and PV+ interneurons, with few NG and VIP+ interneurons. Overall, each IN cluster was

composed of multiple interneuron types. For interneurons in IN8, they were mainly consisted by SST+ and PV+ cell types with a few VIP+ cell type.

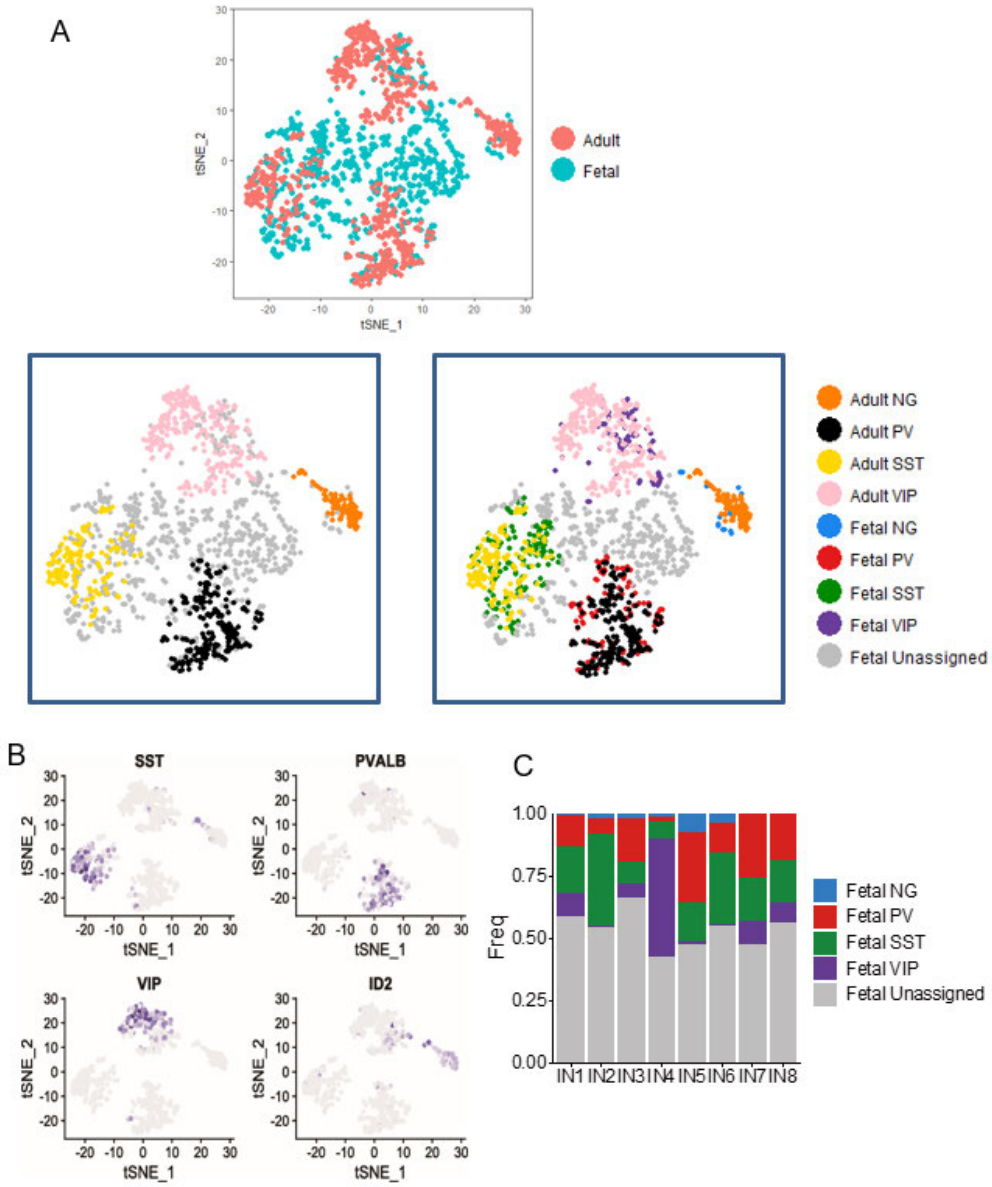


Figure 29: Characterization of interneuron group with enriched expression of ASD risk genes.

(A) Integration of embryonic neurons and adult cortical interneurons in a t-SNE space following CCA. Left, cells are coloured by experimental samples; Middle, cells are coloured by cell type cluster; Right, the assigned embryonic interneuron cell types. (B) Visualization of interneuron diversity among cell types by gradience plots of marker genes. (C) Bar plot depicting the percentage of developmental windows in each interneuron cell cluster in the developing human PFC.

3.5 Discussion

In this chapter, we had set out to address the key question of which cell types specific expressing the ASD risk genes during human brain development. To identify specific cardinal cell classes relevant to ASD during brain development, we grouped the data into six cardinal cell classes: NPCs, ExN, IN, OPC, Astrocyte and Microglia. Our preliminary analysis suggested that 7 of 30 (23.3%) of the *16p11.2* CNV genes and 17 of 83 (20.5%) of the monogenic genes were enriched in one or more cardinal cell classes. And the IN was regarded as a disproportionately vulnerable cell class for ASD since most of differentially expressed ASD risk genes enriched in this class (Figure 30).

In total, 11 genes were identified as significantly enriched in IN, and the function of proteins that encoded by these gene were related with “neurotransmitter” and “ion channel”. In detail, by literature review, at least 5 out of 11 genes encode protein that related with “neurotransmitter”. SLC6A1 protein, belonging to the sodium-neurotransmitter symporter (SNF) family, is a sodium- and chloride-dependent GABA transporter. This protein can terminate the action of GABA transmitter by its high affinity sodium-dependent reuptake into presynaptic terminals (Pramod *et al.*, 2013). RIMS1 protein is a synaptic membrane exocytosis protein. This protein can act as a scaffold protein that regulates neurotransmitter release at the active zone. This protein is also essential for maintaining normal probability of neurotransmitter release and for regulating release during short-term synaptic plasticity (Schoch *et al.*, 2002). GRIA1 protein is a well-known glutamate receptor. As an ionotropic glutamate receptor, it binds with the excitatory neurotransmitter L- glutamate at many synapses in the central nervous system, leads to the opening of the cation channel, and thereby converts the chemical signal to an electrical impulse (Watson, Ho and Greger, 2017). TRIO protein is involved in coordinating actin re-modelling, which is necessary for cell migration and growth. From previous study, we know that this protein can limit the dendrite formation in the

developing hippocampal neurons, without affecting the establishment of axon polarity. Once dendrites are formed, this protein could be involved in the control of synaptic function by regulating the endocytosis of AMPA-selective glutamate receptors (AMPARs) at the excitatory synapses (Ba *et al.*, 2016). Previous experiments also proved that the function of TRIO protein is related with synaptic function and axon guidance within interneuron, but no study described in detail that signalling pathway is regulated by this protein at the inhibitory synapses. DEAF1 protein is an inhibitor of cell proliferation, by arresting cells in the G0 or G1 phase. In recent years, this protein is regarded as a transcription factor which is essential for central nervous system and early embryonic development. Many studies illustrated that this protein affects the 5-HT1A receptor of interneurons which abundantly expressed in post-synapse in rodent models (Philippe, Tristan and Philippe, 2016).

At least 3 out of 11 genes encode protein that related with “ion channel”. In detail, both SCN2A and SCN9A protein belong to sodium channel protein family. Assuming opened or closed conformations in response to the voltage difference across the membrane, these proteins form a sodium-selective channel through which Na+ ions may pass in accordance with their electrochemical gradient. They also can mediate the voltage-dependent sodium ion permeability of excitable membranes (Meisler, O’Brien and Sharkey, 2010). ANK2 protein is required for the coordinated expression of the Na/K ATPase, which is a kind of ion channel related enzyme (Abriel and Kass, 2005). There is one gene related with both “neurotransmitter” and “ion channel”. *NRXN1* gene encodes a cell surface protein involved in cell-cell-interactions, regulates calcium channel activity and plays a role in the regulation of Ca(2+)-triggered neurotransmitter release at synapses and at neuromuscular junctions (Mozhui *et al.*, 2011).

The function of the three proteins left is not clear within interneuron. SEZ6L2 protein, may contribute to specialized endoplasmic reticulum (ER) in neurons. KMT2A protein is a histone methyltransferase that plays an essential role in early development and haematopoiesis. One study reported that neuronal

Kmt2a/Mll1 histone methyltransferase is essential for prefrontal synaptic plasticity in mouse model, but it is not specific to interneuron in that study (Jakovcevski *et al.*, 2015).

Four monogenic genes were enriched in ExN through our analysis. Both *MYT1L* and *BCL11A* genes are transcription factors. BCL11A protein is associated with the chromatin re-modelling complex during brain development. In mouse model, *Bcl11a* gene is an important regulator of terminal neuronal differentiation involved in brain development, and it controls migration of cortical projection neurons through regulation of Sema3c (Wiegreffe *et al.*, 2015). *MYT1L* gene is a transcription factor which has been found only in neuronal tissues. It also plays a key role in neuronal differentiation by specifically repressing expression of non-neuronal genes, as well as negative regulators of neurogenesis (Gao *et al.*, 2011). *CNTNAP2* gene encodes a neuronal transmembrane protein, which is a member of the neurexin superfamily. This protein plays a role in neuron-glia interactions and regulate K+ channels in myelinated axons (Peñagarikano *et al.*, 2011). *LEO1* gene encodes an RNA polymerase-associated protein, and this protein is implicated in regulation of development and maintenance of embryonic stem cell pluripotency (Ding *et al.*, 2015). There is no evidence that *LEO1* gene is specific related with excitatory neurons.

For the CNV genes on *16p11.2*, there were three genes, *PPP4C*, *HIRIP3* and *KIF22*, that were enriched in NPC. KIF22 protein is involved in spindle formation and the movements of chromosomes during mitosis and meiosis. In previous studies, KIF22 protein was implicated in the formation of neural progenitors, as well as maintain cancer cell proliferation in mouse model (Blaker-Lee *et al.*, 2012; Suuberg, 2018). In the human model, we find that KIF22 protein is related with the cell cycle regulation, especially in G2/M phase (Morson *et al.*, 2019). PPP4C protein is a phosphatase enzyme. *Ppp4c* heterozygotes shown growth retardation with decreased survival in Zebrafish model, but there is no study concerning what is the role of *Ppp4c* in NPCs.

HIRIP3 gene encodes a chromatin-related protein, and there is no study concerning what is the role of *HIRIP3* in NPCs as well.

Based on the expression pattern of ASD risk genes among the cardinal cell classes, we find that most of the enriched genes in NPCs are not well studied. Only *Kif22* gene was related with cell proliferation in mouse model. Within ExN, the function of *LEO1* gene in excitatory neurons is still little known. But *Bcl11a, MYT1L* and *CNTNAP2* genes are related with neurogenesis and migration.

IN is a disproportionately vulnerable cardinal cell class for ASD development since most of ASD risk genes are highly expressed in IN. The genes that are significantly enriched in IN can be separated into two categories, “neurotransmitter” and “ion channel”. *SLC6A1, RIMS1, GRIA1* and *TRIO* genes are related with different neurotransmitters and different receptors (Simunovic *et al.*, no date; Devor *et al.*, 2017; Mattison *et al.*, 2018). *SCN2A*, *SCN9A* and *ANK2* genes are related with “ion channel”, especially sodium and calcium channel (Shi *et al.*, 2009; Klassen *et al.*, 2011; Mozhui *et al.*, 2011). This result suggests that multiple molecular mechanisms are involved in the development of ASD, and these mechanisms affect on different cardinal cell classes by type of cellular characteristics.

Notably, the single-cell approach has not only characterized the well-understood cardinal cell classes in the developing human brain, but also investigated the variability of highly expressed genes among novel cell types. Based on the unsupervised clustering, we identified vRG, oRG and IPC in NPCs, DL-like neuron and UL-like neuron in ExN, and several cell clusters in IN. We calculated the significant differential expressed ASD risk genes across different cell types within each neuronal cardinal class. We noted that 6 of the well-established ASD risk genes are enriched in P6, which was regarded as IPCs. The enriched GO terms in P6 are related with axon and dendrite development. It means more ASD risk genes are highly expressed in the neuronal differentiation than the genes highly expressed in the neuronal proliferation. The expression analysis within IN indicated that the developing interneurons in IN8 expressing the highest proportion of ASD risk genes at

relatively high levels. It is not surprising that the unsupervised clustering of embryonic interneurons could not reflect the identity of well-known lineage information, since many lineage maker genes are not expressed in the early developmental stages. The analysis of GO term enrichment only indicated some broadly neuronal mechanisms, as like postsynaptic, synapse and ion channel. We also noticed that *SCN9A* gene is not only differentially expressed across the cardinal cell classes, but also differentially expressed across the cell clusters within interneurons. This gene encodes sodium channel protein, so we assume that the ion channel can not only be used to identify the cardinal cell classes, but also can be used to identify the cell cluster within interneurons. In other words, bioelectric cell properties maybe play an important role in the generation of interneuron diversity (Gelman *et al.*, 2011).

In the analysis about ExN, we noticed that the expression pattern of some well-known maker genes, such as *TBR1*, was different with previous studies, and many maker genes of excitatory neurons at embryonic stage were plotted at low expression levels (Willsey *et al.*, 2013; Parikshak *et al.*, 2013). It means the expression pattern of layer-specific genes within embryonic neurons showed limited correspondence to the expression pattern of these genes within adult cortical neurons. For example, *FOXP2* gene was a widely used laminar marker in adult human cortex but few expressed *FOXP2* in the embryonic neurons in this dataset. It was possible that *ZFRM2, NTSR1, FOXP2, TBR1, OTX1* and *ETV1* maybe good marker genes of layer-specificity in adult human cortex, but all of them were low expressing in this dataset. So, we do not plan to discuss the expression pattern of ASD risk genes between DL-like and UL-like neurons.

Overall, we investigated the differential expressed ASD risk genes through two different steps. Firstly, we did it among cardinal cell classes; then we did it among cell clusters within each cardinal class. Taken together, these findings illustrated that many of the ASD risk genes were differential expressed in major/sub cell types that belong to Progenitor, Interneuron, Astrocyte and Microglia. Previous analysis of these genes indicated that these ASD-affected

genes may play roles in the E/I balance, inhibitory neurogenesis and neuron-glia signalling. Our analysis has revealed gene expression patterns at the single cell level that suggest some cells, most strikingly certain interneurons, may be disproportionally vulnerable to a large number of ASD causing mutations so represent a convergent target for ASD.

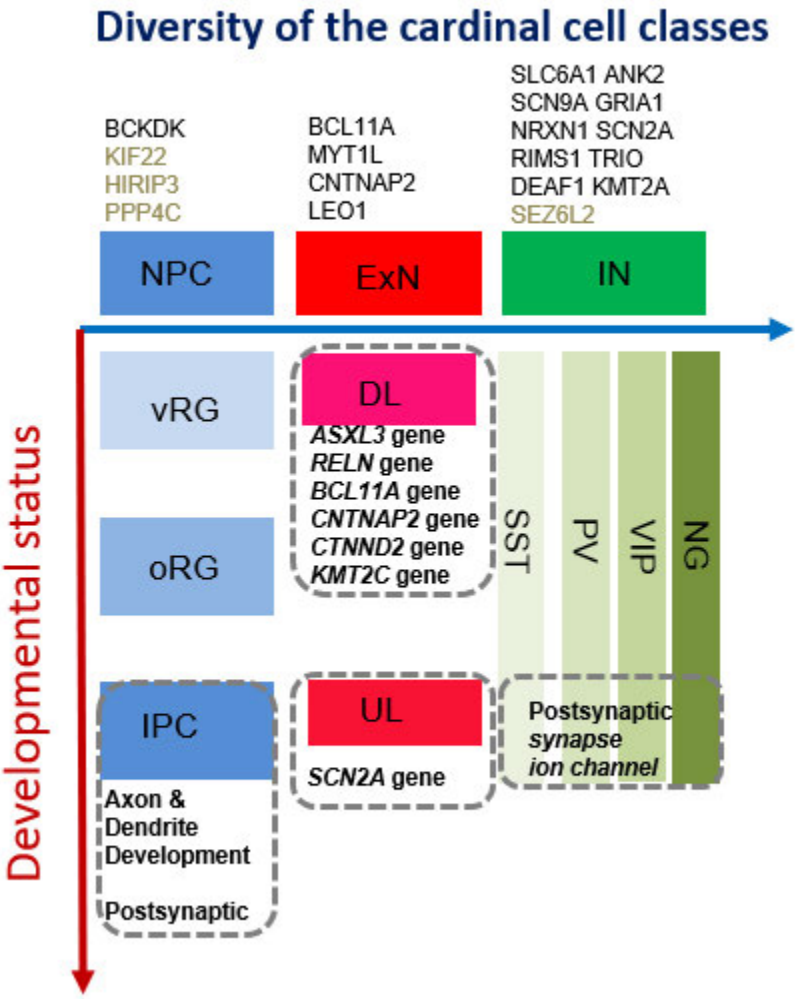


Figure 30: Enrichment of ASD risk genes expression among cell types.

Blue arrow: the different cardinal cell classes within developing human PFC; Red arrow: the different cell types or cell states within each cardinal cell classes.

Charter 4: Investigation of autism gene expression in alternative developing human cortical dataset

4.1 Introduction

Based on the analysis of Zhong's dataset described in Chapter 3, we find that distinct sets of ASD risk genes were enriched in neural progenitor cells, excitatory neurons, interneurons and glia cells. Such enrichments were due to cell subtypes within these cardinal cell classes having significant differences in ASD risk gene expression. Cell-type based gene functional analysis further demonstrated that common biological processes converged on cell subtypes with enriched expression patterns of ASD risk genes. Emergence of this comprehensive scRNA-seq study provided new insights into the cell type-specific molecular pathology of the ASD by comparing their transcriptome expression patterns. But compared to bulk RNA-seq, scRNA-seq methods produces nosier and more variable data, and there were technical issues still awaiting resolution in scRNA-seq analysis.

Since the potential and unavoidable technical issues in scRNA-seq studies, we tried to use another independent dataset to investigate the expression pattern of ASD risk genes as we described in Chapter 1. We looked through the recent published single cell sequencing studies of the human developing cortex (Table 4.1).

4.2 Aim of this chapter

In this Chapter, we aimed to identify the vulnerable cell types underlying the development of ASD during human brain development by checking the expression pattern of ASD risk genes from another published scRNA-seq dataset. Similar as what we did in Chapter 3, we compared the expression pattern of focused ASD risk genes based on two different levels. Firstly, we compared the expression pattern of these genes across the cardinal cell classes within the scRNA-seq dataset. Secondly, based on the unsupervised clustering in each neuronal cell classes in the original paper (progenitor cells, excitatory and inhibitory neurons), we compared the expression pattern of ASD risk genes across the cell clusters in each cardinal cell class, and record what genes were differentially expressed both in Zhong's and Nowakowski's datasets. Finally, we discussed the difference of differentially expressed ASD risk genes between two datasets.

4.3 Summarising available human fetal cortical single cell sequencing datasets

In Chapter 3, we used Zhong's dataset to reveal the expression pattern of ASD risk genes among the various of cell types. Zhong et al. used scRNA-seq (mouth pipette, SMART-seq2, full length) to identify the molecular signatures that mark the diversity of neuronal cell types. The 2,306 cells in this dataset were collected from human developing PFC, and the tissues age of sampling range from GW8 to GW26. Average 2,564 genes were detected per cell in this full-length sequencing dataset, and the authors were able to bring resolution of neuron diversity into their scRNA-seq data by unsupervised clustering of cellular transcriptomic profiles. This enabled the

authors to identify early transcriptional features that instruct the sequence and pace of neuronal differentiation events in the human developing PFC.

Pollen et al. used scRNA-seq (Fluidigm C1 microfluidic platform, SMART-seq, full length) to identify the molecular signatures that mark radial glia cells located in the outer sub-ventricular zone (Pollen *et al.*, 2015). The authors micro-dissected 393 cells in the VZ and OSVZ and used scRNA-seq data from each location to classify and identify distinct RG populations (vRGs and oRGs) in the developing human cortex (GW16-18). Average ~3,000 genes were detected per cell in this full-length sequencing dataset. Their results shed light on the molecular characteristics that establishes the oRG identity in OSVZ, such as the production of trophic factors and extracellular matrix proteins, and the activation of the STAT3 signalling pathway.

Darmanis et al. used scRNA-seq (Fluidigm C1 microfluidic platform, SMART-seq, full length) on 466 cells to capture the cellular complexity of the adult and fetal human brain at a whole transcriptome level (Darmanis *et al.*, 2015). Healthy adult temporal lobe (TL) tissue was obtained from epileptic patients during temporal lobectomy for medically refractory seizures. Fetal human cortical neurons were collected from prenatal brain at the age of GW16–18. Average ~4,000 genes were detected per cell in this full-length sequencing dataset. The authors were able to classify individual cells into all of the major neuronal, glial, and vascular cell types in the brain. And they identified genes that are differentially expressed between fetal and adult neurons and those genes (for example *MKI67*, *PAX6*, *TUBB3* and *DCX* genes) display an expression gradient that reflects the transition between replicating and quiescent fetal neuronal populations. Moreover, they observed the expression of major histocompatibility complex type I (MHCI) genes (for example *TAPBP* and *ERAP1* genes) in a subset of adult neurons, but not fetal neurons.

Fan et al. performed scRNA-seq (cell pellet was collected manually, Single-cell tagged reverse transcription sequencing (STRT-seq), 3’ end) on 4,213

single cells from 21 different regions of the entire human cortex at GW24 and GW25 (Fan *et al.*, 2018). More than 4,000 genes were detected per cell in this full-length sequencing dataset. The authors identified 29 different cell types, which showed different proportions in each region. And they revealed the molecular differences of regional development in the whole human cerebral cortex at the mid-gestational stage.

Nowakowski et al. investigated the transcriptome of single cells (Fluidigm C1 microfluidic platform, SMART-SEQ, full length) of human fetal cortex and medial ganglionic eminence across key stages of prenatal neurogenesis (from GW8 to GW39) (Nowakowski *et al.*, 2017). Average ~2,400 genes were detected per cell in this full-length sequencing dataset. Analysis and clustering of 4,261 cells revealed lineage-dependent trajectories of transcriptional regulators, and that modest transcriptional differences in cortical radial glial stem cells cascade into robust cell-type-dependent differences in neurons.

4.3.1 Comparing available human fetal cortical single cell sequencing datasets

We compared the tissues age of sampling and data quality across these datasets, and we aimed to find a dataset that have a similar sequencing quality report with Zhong's dataset. In other words, we want to find a dataset that including thousands of cells collected from developing human cortex, a wide range of tissue sampling ages and a reasonable number of genes detected per cell.

In detail, Pollen et al. collected cells only from VZ and OSVZ in a narrow range of developmental stages, meaning most of cells in this dataset were likely to be RG cells, and only a small number of ExNs and INs could be collected. So, it was not comparable with Zhong's dataset. Darmanis et al.

collected cells from the whole human developing cortex, but the tissue ages of sampling were limited from GW16 to GW18, and only hundreds of fetal human cortical cells were included in this dataset. This dataset was not comparable with Zhong's dataset. Fan et al. collected more than 4,000 cells from the whole human developing cortex, and more than 4,000 genes were detected per cell. But the question is the tissue ages of sampling limited in GW24 and GW25 in this dataset. Another important different point between Fan's and Zhong's dataset was the cDNA library. Zhong et al. used the full-length sequencing library, but Fan et al used 3' end sequencing library. We cannot judge if the difference of library preparation could affect the detection of ASD risk genes. Nowakowski et al. collected more than 4,000 single cells from multiple human fetal cortical regions, including PFC, across a wide range of key stages of prenatal neurogenesis. They prepared full-length sequencing library, and ~2400 genes were detected per cell. All parameters about Nowakowski's dataset were very similar and comparable with Zhong's dataset. We decided to choose Nowakowski's dataset as the independent dataset to verify the expression pattern of differentially expressed ASD risk genes that we revealed in Zhong's dataset.

Reference	Method	cDNA type	Age	# of cells	Source of cells	Seq depth	Genes /cell
Pollen et al.	C1	Full-length	16-18GW	393	VZ/OSVZ	2.5M	3099
Darmanis et al.	C1	Full-length	16-18GW 21-63y	466	TL/ Cortex	2.8M	~4000
Fan et al.	STRT	3'	24-25GW	4213	22 brain regions	2M	4318
Zhong et al.	Smart-Seq2	Full-length	8-26GW	2306	PFC	?	2654
Nowakowski et al.	C1	Full-length	6-37GW	4261	Cortex	2.2M	2403

Table 3: Table summarizing the published scRNA-seq datasets about human developing cortex.

4.3.2 Overview of the alternative dataset

The published dataset we used in this Chapter was a scRNA-seq dataset stored in the UCSC Cell Browser under the path called “cortex-dev”. Nowakowski et al. used scRNA-seq to identify diverse neuronal subtypes as well as temporally- and spatially- restricted trajectories of neuronal differentiation and maturation across different cortical areas. In this chapter, all data was obtained through the data repositories as described in the original paper and was used without any additional processing. We downloaded the gene expression matrix, and the transcript counts of each cell were normalized to counts per million (CPM), where CPM is the transcript count of each gene divided by the sum of transcript counts of that cell, multiplied by one million. The original classification result of discrete neuronal subtypes in the paper and the cluster interpretation were downloaded from the supplemental materials in the paper. We filtered out the cells that were identified as not from PFC region, leaving 1,125 PFC cells for further analysis. In order to keep comparable with Zhong’s dataset, we combined some subtypes of cells in cardinal classes manually based on the cluster interpretation (Table 4.2).

In details, 1,125 cells were collected from developing human PFC. In order to get the same cardinal cell classes as in Zhong’s dataset, we dropped the Endothelial cells (11 cells), Glycolysis cells (2 cells) and Mural cells (9 cells). Two small groups of cells that identified as “MGE-RG” (1 cells) and “MGE-IPC” (2 cells) were removed as contaminators. 49 cells were dropped since they were not classified as any cell class.

The 1,051 cells left were classified as NPCs, ExNs, INs, OPC, microglia and astrocytes following the original cluster interpretation. For NPCs, we manually combined “IPC-div1” and “IPC-div2” together since both expressed IPC markers, as well as RG markers. “IPC-nEN1”, “IPC-nEN2” and “IPC-nEN3” were manual combined as “IPC_ExN-like” as these cells expressing both IPC and ExN markers. For ExNs, the two groups of ExN (“nEN-early1” and “nEN-

early2") were combined as "ExN_earlyBorn" because these cells were described as the early newborn excitatory neuron in the original cluster interpretation. The "nEN-late" cells were not changed as the late newborn excitatory neuron. Strikingly, the three ExN groups ("EN-PFC1", "EN-PFC2" and "EN-PFC3") were describe as "Early and Late Born Excitatory Neuron PFC", which means these cells cannot be labelled as either deep layer (early born ExNs) nor upper layer (late born ExNs) based on the expression pattern of maker genes. We labelled these groups of cells as "Others". The last ExN group, "EV-V1", were regarded as a contaminator of tissue capture, since these cells were labelled as PFC cells in the original paper. We dropped this group of cells. For INs, we keep the original clustering result but drop the four small cell groups ("IN-STR", "nIN1", "nIN2" and "nIN3") that described as the contaminated interneurons that come from striatum and MGE. After filtering and combination, 202 NPCs, 406 ExNs, 282 interneurons, 21 OPC, 40 microglia and 29 astrocytes were left for the next analysis.

Similar to the analysis in Chapter 3, these 980 cells are labelled as developmental window (W) 1, W2, W3 or W4 based on the tissues age of sampling, and the CPM expression matrix of cells were used to create *Seurat* object followed by log normalization using log (CPM+1). Only protein coding genes that present in at least 0.5% of the cells were used to do differentially expression analysis. We then calculated the differentially expressed ASD risk genes between cell groups using the same methods as in Chapter 3, and recorded what genes were differentially expressed across the cell groups in both Zhong's and Nowakowski's datasets.

Table 4: Table summarizing the clustering result in the original paper and the re-grouped result we used in this Chapter.

	Original result			Re-grouped result	
	cellType	**# of cells**	**Original Cluster Interpretation**	**cellType**	**# of cells**
NPCs	vRG	33	Ventricular Radial Glia	vRG	33
	RG-div1	33	Dividing Radial Glia (G2/M-phase)	RG div1	33
	RG-div2	26	Dividing Radial Glia (S-phase)	RG div2	26
	tRG	25	Truncated Radial Glia	tRG	25
	oRG	21	Outer Radial Glia	oRG	21
	IPC-div1	12	Dividing Intermediate Progenitor Cells RG-like	IPC_RG-like	16
	IPC-div2	4	Intermediate Progenitor Cells RG-like		
	IPC-nEN1	13	Intermediate Progenitor Cells EN-like	IPC_ExN-like	48
	IPC-nEN2	28	Intermediate Progenitor Cells EN-like		
	IPC-nEN3	7	Intermediate Progenitor Cells EN-like		
	Total NPCs	**202**		**Total NPCs**	**202**
ExNs	nEN-early1	6	Newborn Excitatory Neuron - early born	ExN_earlyBorn	146
	nEN-early2	140	Newborn Excitatory Neuron - early born		
	nEN-late	83	Newborn Excitatory Neuron - late born	ExN_lateBorn	83
	EN-PFC1	63	Early Born Deep Layer/subplate Excitatory Neuron PFC	Others	177
	EN-PFC2	69	Early and Late Born Excitatory Neuron PFC		
	EN-PFC3	45	Early and Late Born Excitatory Neuron PFC		
	EN-V1	58	Excitatory Neuron V1	Dropped	
	Total ExN	**464**		**Total ExN**	**406**
INs	IN-CTX-MGE1	89	MGE-derived Ctx inhibitory neuron, Germinal Zone Enriched	IN-CTX-MGE1	89
	IN-CTX-MGE2	42	MGE-derived Ctx inhibitory neuron, Cortical Plate-enriched	IN-CTX-MGE2	42
	IN-CTX-CGE1	77	CGE/LGE-derived inhibitory neurons	IN-CTX-CGE1	77
	IN-CTX-CGE2	74	CGE/LGE-derived inhibitory neurons	IN-CTX-CGE2	74
	IN-STR	7	Striatal neurons	Dropped	
	nIN1	2	MGE newborn neurons		
	nIN2	3	MGE newborn neurons		
	nIN3	1	MGE newborn neurons		
	Total INs	**295**		**Total INs**	**282**
	OPC	21	Oligodendrocyte progenitor cell		21
	Microglia	40	Micrgolia		40
	Astrocyte	29	Astocyte		29
	Total	**1051**		**Total**	**980**

4.4 Results

4.4.1 The cellular heterogeneity among the cardinal cell classes

Based on the cell filtering at Table 4.2, we analysed the expression profiling of protein coding genes among 980 cells. For the expression levels of genes, CPMs were obtained from the original paper. Six cardinal cell classes were revealed in this dataset as the authors' original classification of cell classes: neural progenitor cells (NPCs), excitatory neurons (ExNs), interneurons (INs), oligodendrocyte progenitor cells (OPCs), astrocytes, and microglia (Figure 31A). In order to represent all cells in a two-dimensional space, we used a similar method as we conducted in Zhong's dataset. Firstly, PCA was performed by *RunPCA* function using DEGs across six classes, and the statistically significant principal components (PCs) that drive systematic variation were identified. Then significant PCs identified by jackstraw analysis were used as input to draw two-dimensional coordinates of tSNE space. Finally, visualization of the cardinal cell classes was coloured in t-SNE space and dots indicated individual cells (Figure 31B).

These prefrontal cortical cells were collected from early to late-gestation (GW8 to GW37). The cortical development stages in this dataset were divided into four windows based on milestones including neurogenesis, differentiation, migration and synaptic function, and we dissected the divergent proportions of developmental windows (Ws) across cell classes. Histograms illustrate the relative contribution of Ws to each cardinal cell class in this dataset (Figure 31C). Most of the cells that belong to NPCs and ExNs were captured from W1 and W2, and there were also a few cells captured from W3 within NPCs and ExNs. Cortical interneurons in this dataset were captured equally from W2 and W3. The glia cells, such as OPC and astrocyte, were captured from W3 and W4. Microglia were consisted by a majority of W2 cells and a part of W4 cells. For the neuronal cell classes, the distribution of Ws in this dataset was not the

same as we described in Chapter 3. In Chapter 3, NPCs existed at W1, ExNs were sequentially generated and collected around W2, and interneurons travel tangentially within the marginal and intermediate zones and collected from W2 and W3.

These cell clusters showed distinct cardinal class aggregation and specific gene expression profiles associated with neuronal classification (Figure 31D). A list of well-known cell class markers that we used in Chapter 3 was used to illustrate the classification across six cardinal cell classes (Darmanis *et al.*, 2015; Pollen *et al.*, 2015; Nowakowski *et al.*, 2017). *PAX6*, *HES1* and *VIM* were used as markers to identify NPCs. *NEUROD2*, *NEUROD6* and *RBFOX1* were markers of ExNs. *GAD1*, *GAD2*, *DLX1* and *DLX2* were widely used markers of INs. *OLIG1*, *OLIG2* and *COL20A1* were OPC markers. *GFAP*, *AQP4* and *SLCO1C1* were used as markers to identify the astrocyte. *PTPRC* and *P2RY12* were used as markers of microglia. The expression pattern of these marker genes shown that the cells were correctly identified.

A

cellClasses	# of cells
NPC	202
ExN	406
IN	282
OPC	21
Astrocyte	29
Microglia	40

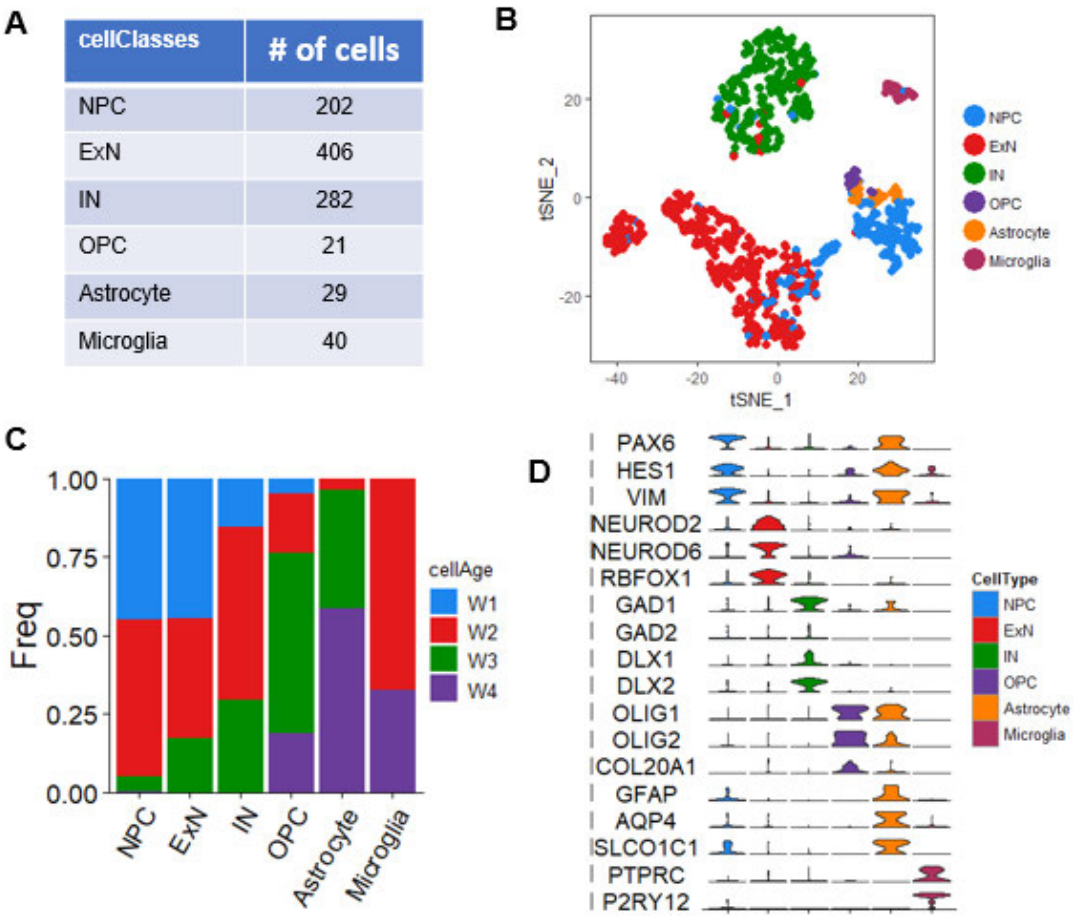


Figure 31: Overview of the scRNA-seq data in Nowakowski's dataset.

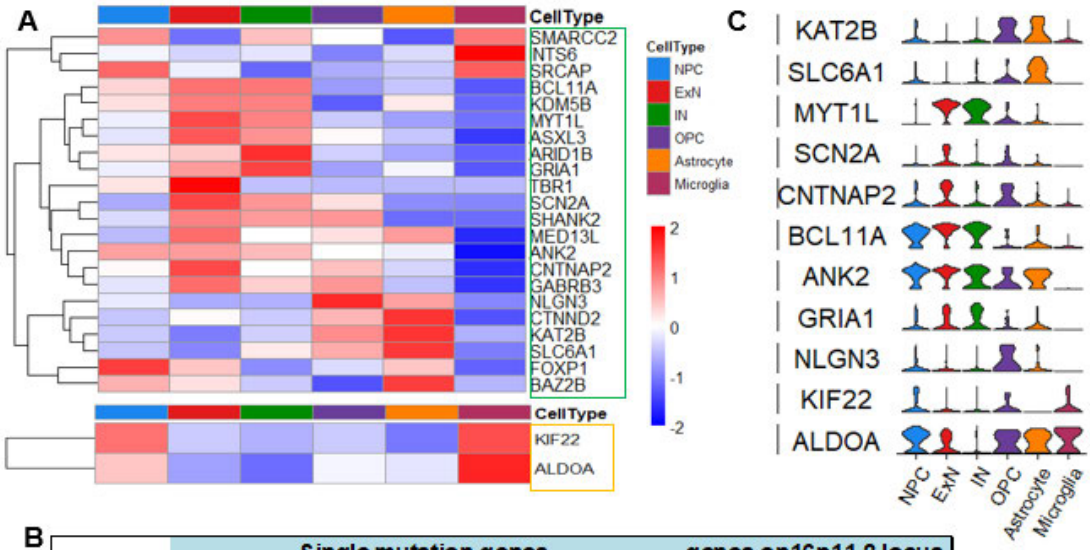
(A) Table summarizing the number of cells in each cardinal cell class. (B) t-SNE plot showing the cardinal cell classes in the dataset. (C) Bar plot depicting the percentage of developmental windows in each cardinal cell classes. (D) Violin plot illustrating the expression pattern of marker genes of six cardinal cell classes. NPC, neural progenitor cells; ExN, excitatory projection neurons; IN, interneurons; OPC, oligodendrocyte progenitor cells.

4.4.1.1 Expression pattern of ASD risk genes among the cardinal cell classes

We revealed the expression pattern of monogenic ASD risk genes and CNV genes on *16p11.2* locus across six cardinal cell classes in the Nowakowski's dataset. (Figure 33 and 34).

We plotted the expression pattern of significant differentially expressed ASD risk genes that identified across the cardinal cell classes (Figure 32A). Twenty-two monogenic genes and two CNV genes on *16p11.2* locus was significant differentially expressed across six cardinal cell classes in Nowakowski's dataset. From the heatmap, we observed most of monogenic genes were enriched in ExNs and INs, and both CNV genes were enriched in NPCs and microglia (Figure 32B).

In detail, the ExN was regarded as a vulnerable cell class for ASD since most of differentially expressed ASD risk genes enriched in this class. Nine monogenic genes and two CNV genes on *16p11.2* locus were identified as significant differentially expressed in both Zhong's and Nowakowski's datasets (Figure 32C). The two genes on *16p11.2* locus, *ALDOA* and *KIF22* genes, were highly expressed in both NPCs and microglia in Nowakowski's dataset. *KIF22* gene was significantly enriched in NPCs, and *ALDOA* gene was significantly enriched in microglia. The expression pattern of *ALDOA* gene was slightly different with the pattern in Zhong's dataset. In Zhong's dataset, *ALDOA* gene was significantly enriched in astrocyte, and expressed in part of NPCs and microglia. The roles of these significant differentially expressed genes and the possible reason of the difference between two datasets will be discussed later in this Chapter.



B

	Single mutation genes	genes on16p11.2 locus
NPC	*FOXP1, SRCAP, SMARCC2, BAZ2B*	*KIF22, ALDOA*
ExN	*TBR1, MYT1L, ASXL3, CNTNAP2, SCN2A, BCL11A, SHANK2, GABRB3, ANK2, KDM5B, CTNND2, MED13L*	
IN	*SLC6A1, ARID1B, GRIA1*	
OPC	*NLGN3, KAT2B*	
Astrocyte	*SLC6A1, KAT2B*	
Microglia	*INTS6*	

Figure 32: Expression pattern of ASD risk genes among cardinal cell classes within Nowakowski's dataset.

(A) Heatmap illustrating the expression pattern of the significant differentially expressed ASD risk genes across cardinal cell classes in Nowakowski's dataset (Wilcox test, adjust $p < 0.05$, log (fold change) > 0.3). Green box: single mutation ASD risk genes; yellow box: CNV genes on *16p11.2* locus. (B) Table summarizing enrichment of the significant differentially expressed ASD risk genes in Nowakowski's dataset. (C) Violin plot illustrating the expression pattern of differentially expressed ASD risk genes across cardinal cell classes both in Zhong's and Nowakowski's dataset.

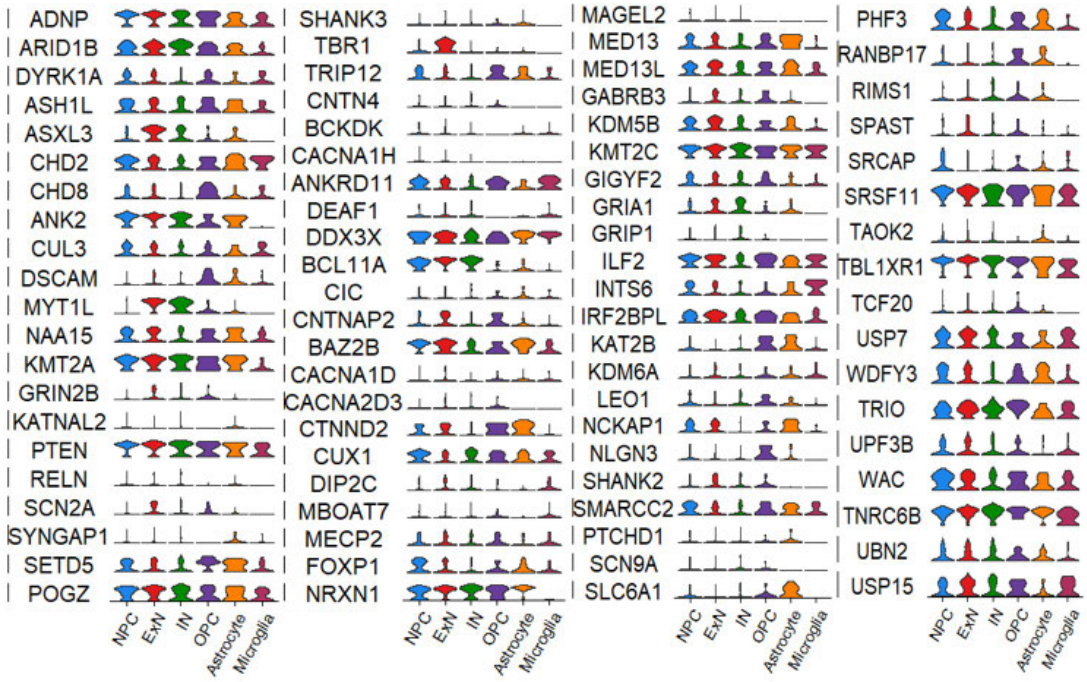


Figure 33: Violin plot illustrating the expression pattern of single mutation ASD risk genes among six cardinal cell classes.

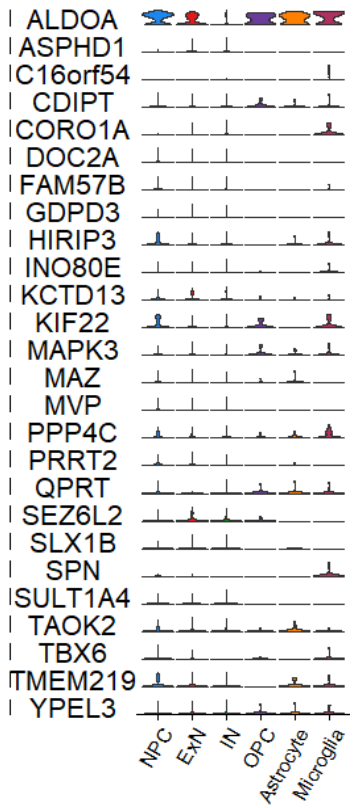


Figure 34: Violin plot illustrating the expression pattern of CNV genes on *16p11.2* locus among six cardinal cell classes.

4.4.2 Analysis of cell types in each cardinal cell class

In each cardinal cell class, there are several distinct cell subpopulations. To further reveal the expression pattern of ASD risk genes across the potential subpopulations within each cardinal cell class, we performed a more detailed analysis within each cardinal cell class. As we described previously, we filtered out the cells that not collected from PFC region, and manually combined a few small groups of cells together to get broader and meaningful cell subpopulations. *t*-SNE plot showing the cell subpopulations in each cardinal cell classes in the dataset. OPC, astrocytes, and microglia are not further analysed.

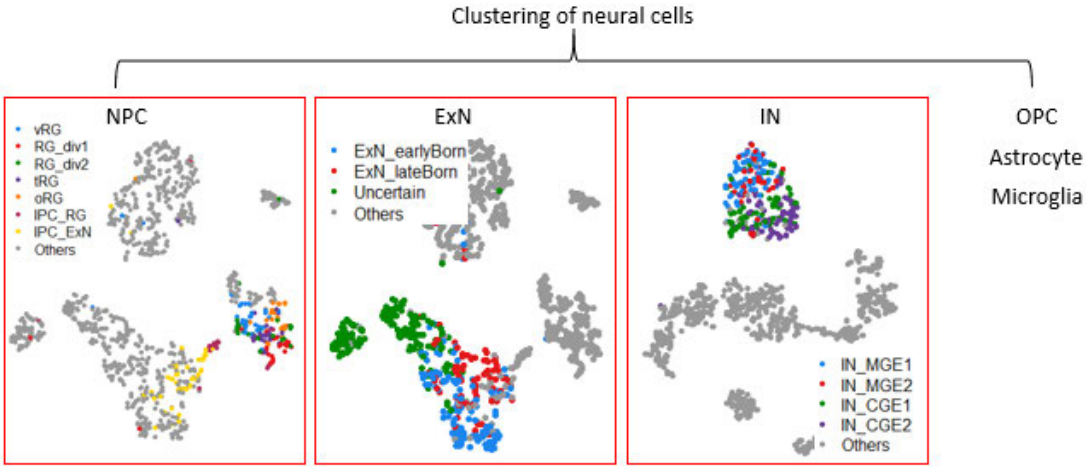


Figure 35: Distinct cell types in cardinal cell classes were represent in the t-SNE space.

Cell type classification was obtained directly through the original paper. OPC, astrocyte and microglia are not further clustered. Dots, individual cells; Colour, cell types.

4.4.3 Expression pattern of ASD risk genes in neural progenitor cells of human fetal cortex

We manually re-organised the clustering result in the original paper. Histograms illustrated the contribution of developmental windows to each cell cluster was very similar to each other (Figure 36A). We used a set of well-known maker genes of progenitor cell types to check the identification of our re-organised clusters (Figure 36B). All cells in six clusters were marked by expression of progenitor markers *PAX6*, *VIM* and *HES1*. *CRYAB*, *NFATC2* and *GPX3* were marker genes of tRG cells that identified in the original paper, but these genes were also expressed in many oRG cells. *HOPX*, *TNC* and *MOXD1*, which have been identified as markers of oRG were high expressed in oRG cluster, while *EOMES*, *PPP1R17*, *NHLH1, NHLH2* and *RBFOX1* was expressed in IPC cluster, which suggests that the cells in the cluster were IPCs or early state of ExNs. Notably, all oRG markers (*HOPX*, *TNC* and *MOXD1*)

and some maker genes of tRGs (e.g., *NFATC2* and *GPX3*) were enriched in RG-div1 and RG-div2 cells. This suggests that RG-div1 and RG-div2 may contain both tRG and oRG cells. Both RGs markers (*PAX6*, *VIM* and *MOXD1*) and IPC markers (*EOMES* and *PPP1R17*) were high expressing in IPC_RG cluster, meaning these cells under a transition state between RGs and IPCs. The cells in IPC_ExN cluster were low expressing RGs markers, but high expressing both IPCs markers (*EOMES, NHLH1* and *PPP1R17*) and ExNs markers (*TBR1*). The expression pattern indicated these cells were differentiating from NPCs to ExNs.

We examined the expression pattern of the monogenic ASD risk genes and genes on *16p11.2* locus among seven NPC clusters in the developing human brain (Figure 37 and 38). We have identified a set of novel marker genes for these progenitor cluster by differential expression analysis (Wilcox test, adjust $p < 0.05$, log (fold change) > 0.3) (Figure 36C). In the NPCs of Zhong's dataset, we find that five monogenic genes (*MYT1L, ASXL3, CNTNAP2, IRF2BPL* and *CUL3*) as well as one gene on *16p11.2* locus (*SEZ6L2*) were included in the DEGs across six clusters. But only two of them (*MYT1L* and *CNTNAP2*) were significantly enriched in IPC_ExN cluster in this dataset (Figure 36D). In this dataset, besides *MYT1L* and *CNTNAP2* genes, *TBR1* gene was enriched in IPC_ExN, and *KIF22* gene was enriched in both RG_div1 and IPC_RG clusters.

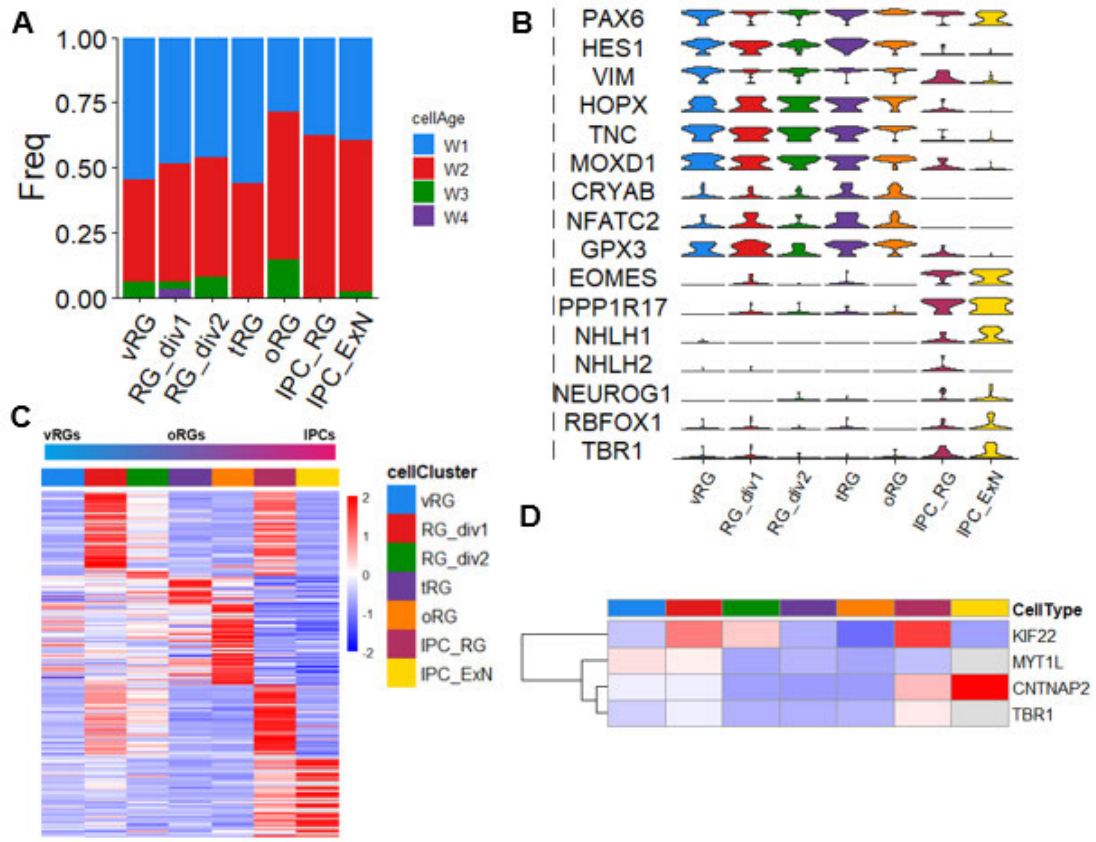


Figure 36: Diversity of cortical progenitor cell types in the human fetal cortex in Nowakowski's dataset.

(A) Bar plot depicting the percentage of developmental windows in each cell type. (B) Violin plot illustrating the expression pattern of marker genes of six cardinal cell classes. (C) Heatmap illustrating the expression pattern of significant differentially expressed genes across cell clusters in NPCs (Wilcox test, adjust p < 0.05, log (fold change) > 0.3). The ventral radial glia cells (vRGs), outer radial glia cells (oRGs) and intermediate progenitor cells (IPCs) are defined by the expression of known markers as listed in B. (D) Heatmap illustrating the expression pattern of differentially expressed ASD risk genes that identified across NPC groups. Green box: single mutation ASD risk genes; yellow box: CNV genes on *16p11.2* locus.

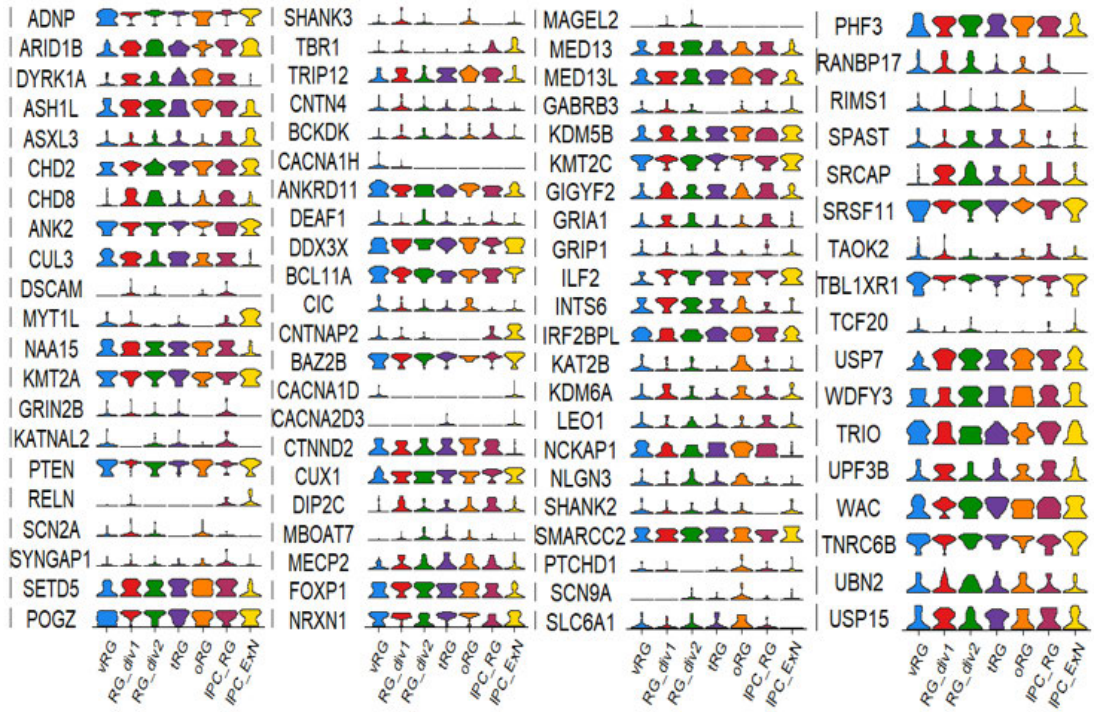


Figure 37: Violin plot illustrating the expression pattern of single mutation ASD risk genes among seven NPC clusters.

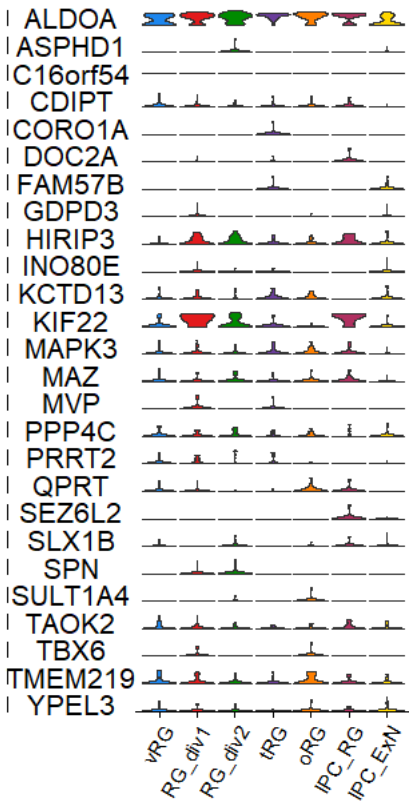


Figure 38: Violin plot illustrating the expression pattern of CNV genes on *16p11.2* locus among seven NPC clusters.

4.4.4 Diversity of excitatory neurons in human developing PFC

Before comparing the expression levels of ASD risk genes across ExNs cell subpopulations, we combined the "nEN-early1" and "nEN-early2" as one "ExN_earlyBorn" cell group since both clusters described as early born excitatory neuron. "ExN-PFC 1/2/3" was not labelled as either "ExN_earlyBorn" or "ExN_lateBorn" since the original description ("Early and Late Born excitatory neuron") was not clear indicate what kind of ExN they belong to.

Histograms illustrated the clustering as they were not biased by the contribution of developmental windows to each cell cluster was very similar to each other (Figure 39A). Cells in ExN_earlyBorn cluster came from W1 and

W2, and the majority of cells in ExN_lateBorn cluster came from W2. Lots of W1 cells and a few of W2 and W3 cells were labelled as “Others”.

As described in Chapter 1, excitatory neurons are generated sequentially in an inside-out order from progenitors residing in the VZ and SVZ. This results in the sequential generation of early deep layer (DL) and late upper layer (UL) neurons, as early-born neurons settle in deep layers of the cortex, whereas late-born neurons populate the upper layers.

We could regard most of neurons in ExN_earlyBorn were likely to be DL-like excitatory neurons, and neurons in ExN_lateBorn likely to be UL-like excitatory neurons. Some maker genes were used to check the layer-specificity of these clusters (Figure 39B). Some deep layer markers, such as *TBR1, SOX5,* and *BCL11B* were high expressed in ExN_earlyBorn cells. Surprisingly, the upper layer markers, such as *CUX1, CUX2* and *SATB2*, were not enriched in ExN_lateBorn cells. Other upper layer markers, such as *UNC5D, RORB, WFS1* and *RELN*, were not enriched in any cluster as well. This means the deep layer markers among embryonic neurons can partly define the layer-specificity of DL-like neurons (ExN_earlyBorn), but marker genes of upper layer showed limited ability to distinguish the UL-like neurons (ExN_lateBorn).

A set of novel marker genes were identified for these excitatory neuron clusters by differential expression analysis (Figure 39C). We examined the expression pattern of the monogenic ASD risk genes and genes on *16p11.2* locus among three excitatory neuron clusters in the developing human brain (Figure 40 and 41). Eleven monogenic genes were included in the significant DEGs across three clusters (Figure 39D). *CUX1* gene was enriched in ExN_earlyBorn, and all other ten genes were enriched in the cluster called “Others”.

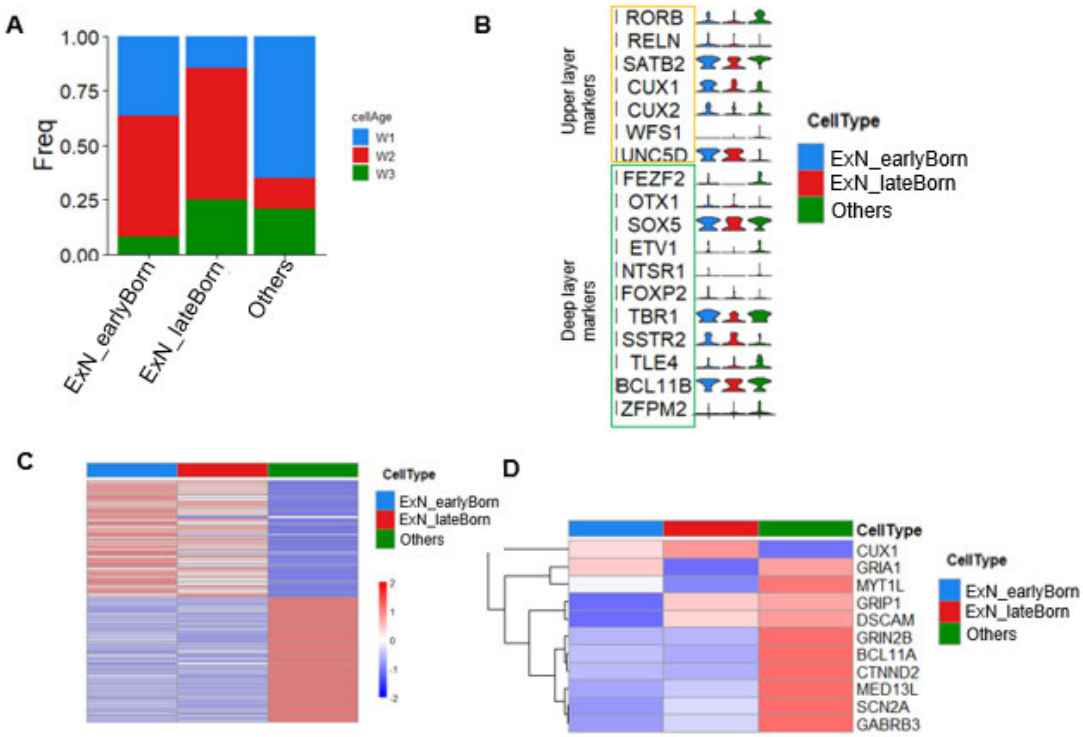


Figure 39: Unsupervised clustering of excitatory neurons in the human fetal cortex in Nowakowski's dataset.

(A) Bar plot depicting the percentage of developmental windows in each cluster. (B) Violin plot illustrating the expression pattern of marker genes between deep layer and upper layer. Yellow box: upper layer maker genes; Green box: deep layer maker genes. (C) Heatmap illustrating the expression pattern of differentially expressed genes across cell clusters within excitatory neurons. (D) Heatmap illustrating the expression pattern of ASD-DEGs across cell clusters. classes in Nowakowski's dataset (Wilcox test, adjust $p < 0.05$, log (fold change) > 0.3).

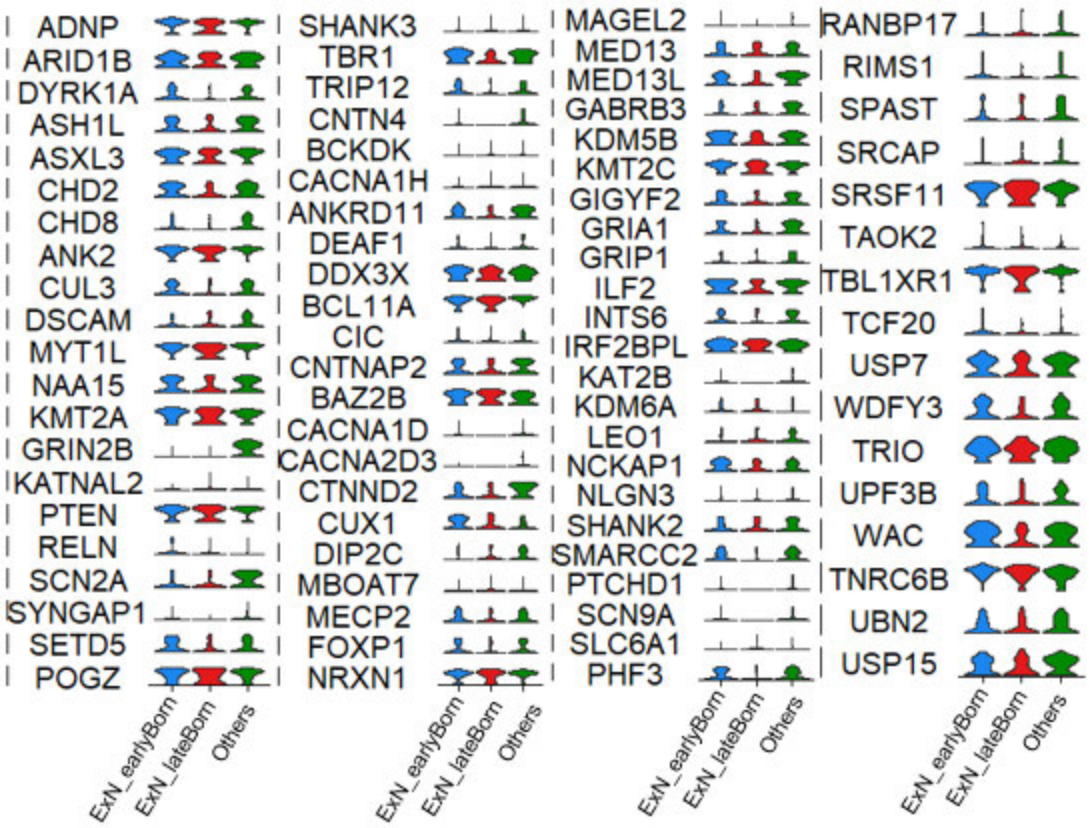


Figure 40: Violin plot illustrating the expression pattern of single mutation ASD risk genes among three ExN clusters.

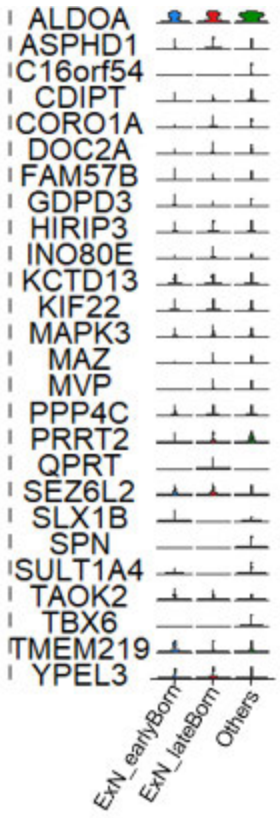


Figure 41: Violin plot illustrating the expression pattern of CNV genes on *16p11.2* locus among three ExN clusters.

4.4.5 Diversity of interneurons in human developing PFC

Four interneuron clusters were identified based on their transcriptional profiling and labelled as IN_MGE1, IN_MGE2, IN_CGE1 and IN_CGE2 in the original paper (Figure 42, IN). Histograms illustrate the relative contribution of Ws to each interneuron cluster (Figure 42A). The majority of interneurons were captured from W1 and W2. IN_MGE2 were mainly consisted by W3 cells with a few W1 and W2 cells. The mixture of developmental windows in the clusters indicated that the clustering was not affected by the sampling time.

The same as what we did in the analysis of interneurons in Zhong's dataset, we used a set of well-known maker genes of interneuron cell types to define

cell identities of the cells in the four clusters. There were fourteen marker genes listed in the violin plot (Figure 42B). The interneurons in IN_MGE1 and IN_MGE2 highly expressed the marker genes of MGE-derived interneurons (*LHX6* and *SOX6*). *SST, TAC1* and *SLIT2* genes regulate the development of MGE-derived subtypes of cortical interneurons, respectively. For the expression pattern of *SST* and *TAC1 genes*, we noticed that they were not differentially expression between IN_MGE1 and IN_MGE2. However, *SLIT2* gene was higher expressed in IN_MGE2 than IN_MGE1, suggesting IN_MGE2 may include some *SLIT2*+ interneuron.

Marker genes of CGE-derived interneurons (*NR2F2, SP8* and *PROX1* genes) were only expressed in IN_CGE1 and IN_CGE2, but not in IN_MGE1 or IN_MGE2. Interesting, the expression levels of these marker genes in the IN_CGE2 cells were much higher than the cells in IN_CGE1. *VIP, CCK* and *ID2* genes were usually regarded as marker genes of VIP+, CCK+, ID2+ cortical interneurons which migrated from CGE region, respectively. The expression levels of *VIP* and *CCK* genes were very low in the two CGE-related clusters. However, *ID2* gene was widely expressed not only in CGE-related clusters, but also in MGE-related clusters. It was very similar with the expression pattern of *ID2* gene in Zhong's dataset that equally expressed across all the interneuron clusters. We did not know whether *ID2* genes are not a good marker for the CGE-specific interneuron or if it was caused by any unknown bias of clustering.

CALB2, RELN and *NPY* genes were marker genes of specific interneuron cell types, and these interneurons were migrated from both MGE and CGE regions. Both *RELN* and *NPY* genes were very few expressed among interneurons in this dataset, but *CALB2* gene was highly expressed in IN_CGE2 cells. It means *CALB2*+ interneurons were grouped in the IN_CGE2 cluster.

We had identified novel marker genes for these interneuron clusters by differential expression analysis (Figure 42C). We examined if any ASD risk genes were significant differentially expressed across the four clusters (Figure

43 and 44). And we found only two genes were significant differentially expressed (Figure 42D). *BCL11A* gene was significant enriched in IN_MGE1 cluster, and *GRIN2B* gene was significant enriched in IN_CGE2 cluster.

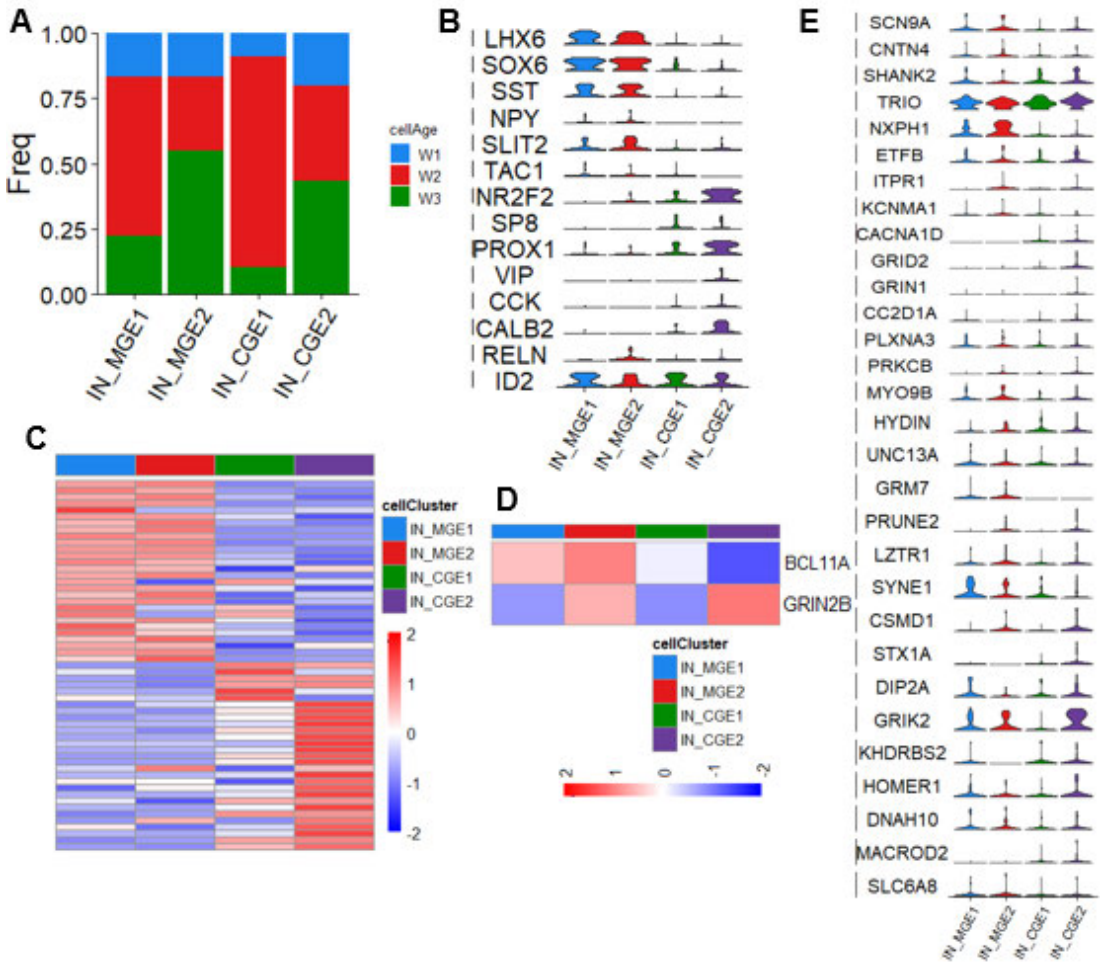


Figure 42: Diversity of interneurons in human developing PFC.

(A) Bar plot depicting the percentage of developmental windows in each cell cluster. (B) Violin plot illustrating the expression pattern of marker genes of six cardinal cell classes. (C) Heatmap illustrating the expression pattern of differentially expressed genes across cell clusters in INs. (D) Heatmap illustrating the expression pattern of significant differentially expressed ASD risk genes across cell clusters in INs. (Wilcox test, adjust $p < 0.05$, log (fold change) > 0.3). (E) Violin plot illustrating the expression pattern of differentially expressed ASD risk genes that identified within Zhong's NPCs.

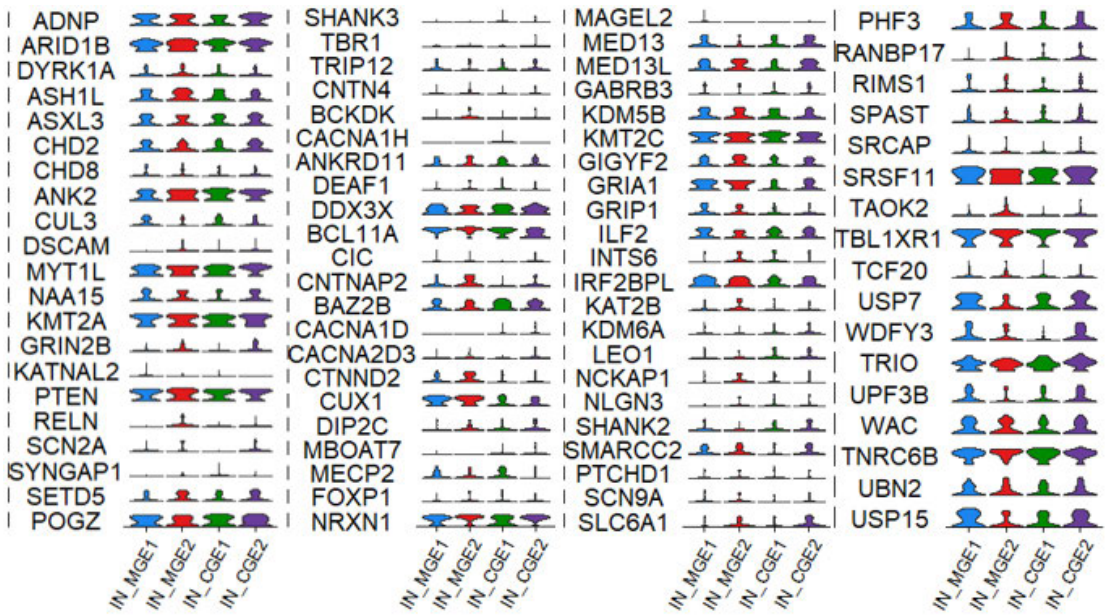


Figure 43: Violin plot illustrating the expression pattern of single mutation ASD risk genes among four IN clusters.

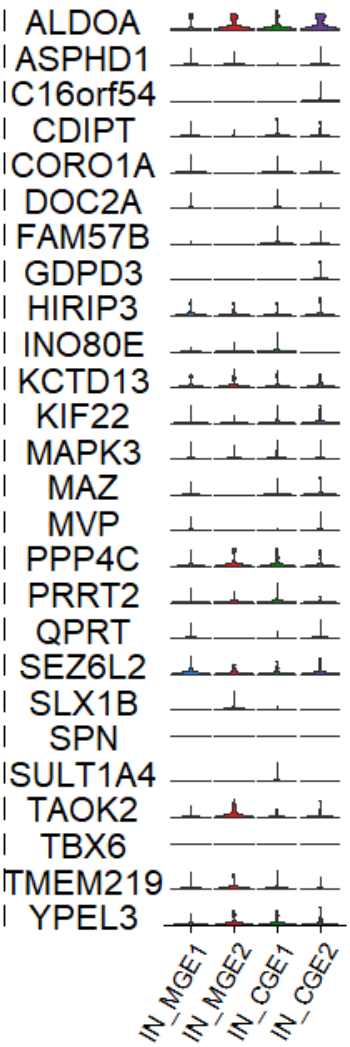


Figure 44: Violin plot illustrating the expression pattern of CNV genes on *16p11.2* locus among four IN clusters.

4.5 Discussion

4.5.1 Differentially expression pattern of ASD risk genes

In this chapter, we checked the expressing the ASD risk genes in an alternative human developing cortical dataset. Six cardinal cell classes were grouped in this dataset: NPCs, ExN, IN, OPC, Astrocyte and Microglia. Our preliminary analysis suggested that 2 of 30 (6.7%) of the *16p11.2* CNV genes and 22 of 83 (26.5%) of the monogenic genes were enriched in one or more cardinal cell classes (Figure 45). Totally 12 genes were identified as significant enriched in ExNs. Most of these genes were also identified as DEGs in Zhong's dataset except *SHANK2, GABRB3, MED13L* and *KDM5B* genes.

By literature review, the other two genes, *SHANK2 and GABRB3*, encoded protein that related with "Synaptic regulation". In detail, SHANK2 is an adapter protein in the postsynaptic density of excitatory synapses that interconnects receptors of the postsynaptic membrane including NMDA-type and metabotropic glutamate receptors. In the animal model experiment, the male mice lacking Shank2 in excitatory neurons and GABAergic inhibitory neurons in the hippocampus and striatum show social interaction deficits (Kim *et al.*, 2018). GABRB3 protein was a component of the receptor for the GABA neurotransmitter, the major in the vertebrate brain, and this protein also function as ligand-gated chloride channel (Mullins, Chung and Rees, 2010). The function of *MED13L* and *KDM5B* genes encoded protein during human cortical neurodevelopment were not clear.

There were six genes that were identified as significant enriched in NPCs. *KIF22* and *ALODA* genes were also enriched in Zhong's dataset. For the other four genes, FOXP1 protein affects embryonic NPCs differentiation by modulating Notch signalling in the developing neocortex (Braccioli *et al.*, 2017). In mouse cortical neurodevelopment, *Baz2b* gene was an IPC marker, it was directly repressed by *Tbr2* and *Tbr1* genes (Elsen *et al.*, 2018). But there

was no detailed explanation about its function in human cortical development. SMARCC2 protein was an intrinsic factor of glial radial cells and plays a crucial role in embryogenesis and corticogenesis, determining the mammalian body and cortical size (Machol *et al.*, 2019). The function of *SRCAP* gene in human NPCs was not clear. We noted that three of the well-established ASD risk genes (*MYT1L*, *CNTNAP2* and *TBR1*) are enriched in IPC_ExN cell cluster. It was not a surprise to us since *MYT1L* gene plays a key role in neuronal differentiation, *CNTNAP2* is a marker gene of DL-like ExNs and *TBR1* is a marker gene of ExNs.

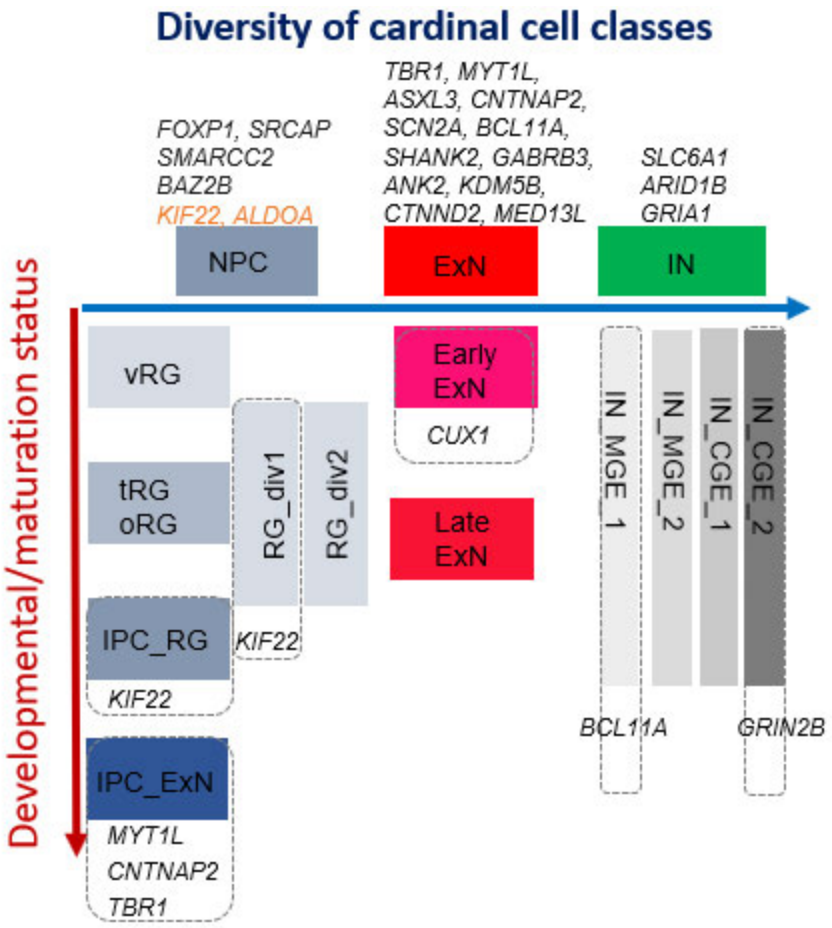


Figure 45: Enrichment of ASD risk genes expression among cell types.

Blue arrow: the different cardinal cell classes within developing human PFC; Red arrow: the different cell types or cell states within each cardinal cell classes. Enriched Monogenic genes were labelled in black italics and genes on *16p11.2* locus was labelled in red italics.

4.5.2 Comparison of the cardinal cell classes between two datasets

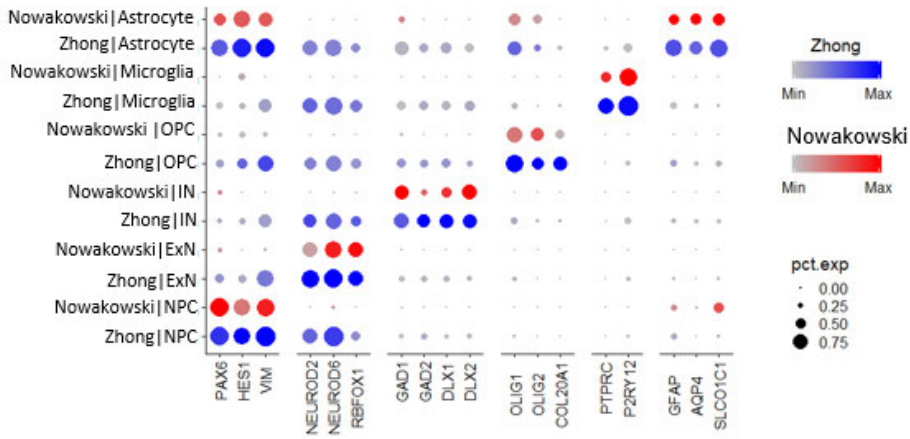
We noticed that the expression pattern of ASD risk genes are different between Nowakowski's and Zhong's datasets. For example, *ASXL3*, *BCL11A* and *CTNND2* genes were only enriched in DL-like ExNs in Zhong's dataset, but these genes were identified as enriched in the whole ExNs in Nowakowski's dataset. *ANK2* and *SCN2A* genes were significant enriched in INs in Zhong's dataset but enriched in the ExNs in Nowakowski's dataset. There were also four novel significant differentially expressed ASD risk genes that enriched in ExNs in Nowakowski's dataset. The first one was *TBR1* gene. This gene was a well-known maker gene of ExNs, but in Zhong's dataset, the expression level of this gene was very low among the ExNs.

In the analysis about NPCs, *KIF22* gene was the only one that significant enriched in NPCs in both Zhong's and Nowakowski's datasets. In the analysis of Nowakowski's dataset, we found that the expression of *KIF22* gene enriched in RG_div1 and IPC_RGs. RG_div1 was indicating the cells in G2/M phase. This pattern proved our previous experiment that KIF22 protein was implicated in the formation of neural progenitors, as well as control G2/M phase of cell cycle in human RGs (Morson *et al.*, 2019).

In order to illustrate the similarity of cardinal cell classes between two datasets, we compared the expression levels and the percentage of expression of the marker genes between the two datasets. It provided an intuitive way to visualize how gene expression changes across different cardinal cell classes (Figure 46A). The cells were split into two groups of colours based on the different datasets. Blue indicated the data from the Zhong's dataset, and red means the data from the Nowakowski's dataset. The size of the dot encodes the percentage of cells within a class, while the colour encodes the average expression level of genes (blue and red are high). This split dot plot showed that the expression pattern of these marker genes is consistent between two datasets.

Cross-datasets validation between the emerging cardinal cell classes in the Zhong's data and the Nowakowski's data confirmed the robustness of these annotations of clustering (Figure 46B). The AUROC score of MetaNeighbor across classes between two datasets provided strong evidence for the specification of cardinal cell classes: all six cardinal classes of cortical cells were identified with strong robustness among fetal PFC cells (red, Figure 46B), which exhibit unique patterns of marker gene expression (Figure 46A). Comparison between the two independent datasets validated all the annotated cardinal cell classes and identification of robust cardinal cell classes allows inference of cellular properties from larger number of cells.

A



B

	Zhong|NPC	Zhong|ExN	Zhong|IN	Zhong|OPC	Zhong|Astrocyte	Zhong|Microglia
Nowakowski|NPC	0.963486013	0.436156378	0.114334971	0.726828145	0.814636636	0.074391642
Nowakowski|ExN	0.386896414	0.929837265	0.407050847	0.302163208	0.284672492	0.32701172
Nowakowski|IN	0.153452561	0.310477831	0.945821248	0.296718595	0.244776489	0.391432856
Nowakowski|OPC	0.566737629	0.358689016	0.451603862	0.963563712	0.797348228	0.276953636
Nowakowski|Astrocyte	0.687154307	0.199668624	0.410707433	0.915651438	0.97735729	0.384930539
Nowakowski|Microglia	0.240016551	0.177343014	0.386232398	0.344810749	0.24365507	0.999363516

Figure 46: Comparative transcriptional analysis between Zhong's and Nowakowski's dataset.

(A) Split dots plot showing the fraction of cells in each cluster expressing a given marker (dot size) and the level of marker gene expression (dot intensity) for marker genes known to exhibit preferential expression in distinct cell classes. Blue, the Zhong's dataset; Red, the Nowakowski's dataset. In the colour bar, grey corresponds to low expression level; blue and red correspond to high expression level. The size of dot indicating the fraction of cells in each cluster expressing a given marker. (B) AUROC scores in the table indicating the correlation between the six cardinal cell classes in Zhong's and Nowakowski's datasets. A mean AUROC score of 0.9 or above typically suggests a reciprocal correlation. The correlation between the same cell cardinal class from two datasets is coloured as red.

4.5.3 Comparison of the sampling ages and sequencing depth between two datasets

The transcriptional profiles of cardinal cell classes were very similar between two datasets, but the expression pattern of ASD risk genes are different. We noticed that, in Zhong's dataset, IN was regarded as a vulnerable cell class for ASD since most of differentially expressed ASD risk genes enriched in this class. But in Nowakowski's dataset, only three genes (*SLC6A1*, *ARID1B* and *GRIA1* genes) were significant enriched in INs. The cortical INs were tangential migrated from GE regions, and the amount of cortical INs was strongly related with the developmental stages. So the number of cells captured and the developmental stages of sampling in a scRNA-seq study could strongly affect the cell type identification of INs. We compared the number of cells captured and the distribution of developmental stages of cell sampling between two datasets (Figure 47). It was notable that in Zhong's dataset, 2,306 cells were collected. But only 980 cells were captured in Nowakowski's dataset (Figure 47A). There were a few cells that collected from W4, but the cells captured from some early developmental stages, such as GW8/9/10/12, could only be found in Zhong's dataset (Figure 47A and B). The contribution of two datasets to each cardinal cell class shown that in every cell class, the number of cells in Zhong's dataset were about twice as high as the number of cells in Nowakowski's dataset (Figure 47C). In details, we compared the distribution of number of cells based on the three cardinal classes of neuronal cells (Figure 48). In NPCs, we found that most of the NPCs in Zhong's dataset were captured from GW9, GW10 and GW16. The NPCs in Nowakowski's dataset were captured from a range of developmental stages from GW15 to GW17. There was a slight difference on the developmental stages of NPCs sampling between two datasets, but we considered they were comparable. For ExNs and INs, there were huge differences between two datasets. In Zhong's dataset, the majority of ExNs were come from one single developmental stage GW16, and a few ExNs were captured from GW23 and Gw26. As a comparison, the ExNs in Nowakowski's had a uniform distribution

from GW13 to Gw24. For INs, most INs in Zhong's dataset were come from two developmental stages, GW23 and GW26. But the INs in Nowakowski's dataset had a uniform distribution from GW15 to Gw24. These distributions could help us explain why we got different conclusion that most of ASD risk genes enriched in ExNs in Nowakowski's dataset but most of ASD risk genes enriched in INs in Zhong's dataset. Since most of the ExNs and INs in Nowakowski's dataset were collected from GW13 to Gw24, there were few maturing interneurons to be found since they need to do tangential migration from GEs to cortex. But at these stages, more maturing ExNs may be found in cortex. So the enrichment of ASD risk genes in ExNs in Nowakowski's dataset may not indicate the expression levels of ASD risk genes were higher in ExNs than NPCs or INs, but mean the expression levels of ASD risk genes were higher in more maturing neurons than progenitors and non-maturing neurons. Compared to Nowakowski's dataset, the enrichment of ASD risk genes was found in INs in Zhong's dataset. This was a more reliable result since the INs at GW26 were more mature, and the ExNs at GW16, GW23 and Gw26 were more likely to be maturing ExNs. So the comparison between the maturing cortical ExNs and the maturing cortical INs was more reasonable.

Lastly, we tried to have a look on the technical biases between two datasets. We found that average ~2,400 genes were detected per cell in the whole Nowakowski's dataset, and average ~2,600 genes were detected per cell in Zhong's dataset. Since we only select the PFC cells in Nowakowski's dataset, we compared the number of genes were detected per cell in PFC region only across developmental stages in each cardinal neuronal cell class (Figure 49). In every cardinal neuronal cell class (NPCs, ExNs, and INs), the number of genes detected per cell in Zhong's dataset were much higher than the number in Nowakowski's dataset. We noticed that most genes on *16p11.2* locus were low expressed in Nowakowski's dataset except *KIF22, HIRIP3* and *ALODA* genes, but many of these genes were normal expressed in Zhong's dataset (Figure 14). For monogenic genes, *SCN9A* gene was significant enriched in a small group of INs in Zhong's dataset, but its expression level was very low in Nowakowski's dataset. We did not have the original sequencing files and

quality check report of two datasets, but the difference in the number of detected genes between two datasets could give us a hypothesis that the number of genes detected per cell may be effected by the sequencing depth or sample quality in two datasets, and the difference between two datasets could lead to the different expression pattern of the of ASD risk genes in two datasets.

A

Zhong's		Nowakowski's		
GW	# of cells	GW	# of cells	Ws
GW08	23			
GW09	88			
GW10	184			
GW12	85			W1
GW13	24	GW13	34	
		GW13.5	35	
		GW15	200	
		GW15.3	47	
GW16	776	GW16	41	
		GW16.2	60	
		GW16.7	124	
		GW17	42	W2
		GW18.5	14	
GW19	120	GW19	62	
		GW19.3	54	
		GW19.5	47	
		GW21	23	
		GW22	36	
GW23	307			W3
		GW23.5	97	
		GW24	29	
GW26	699			
		GW34	26	W4
		GW39	9	
Total #	2306	Total #	980	

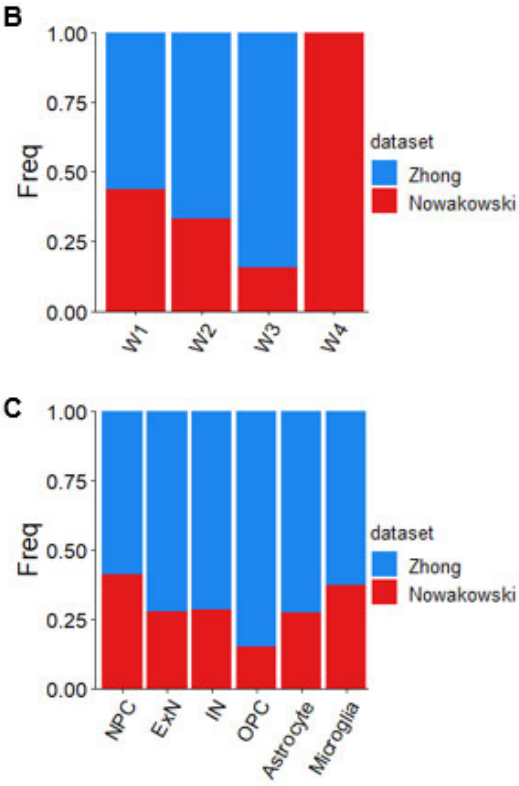


Figure 47: Comparison of the number of cells and the distribution of cell sampling between two datasets.

(A) Table summarizing the number of cells captured and the distribution of developmental stages of cell sampling between two datasets. (B) Bar plot showing the distribution of Ws between two datasets. (C) Bar plot showing the contribution of two datasets to each cardinal cell class.

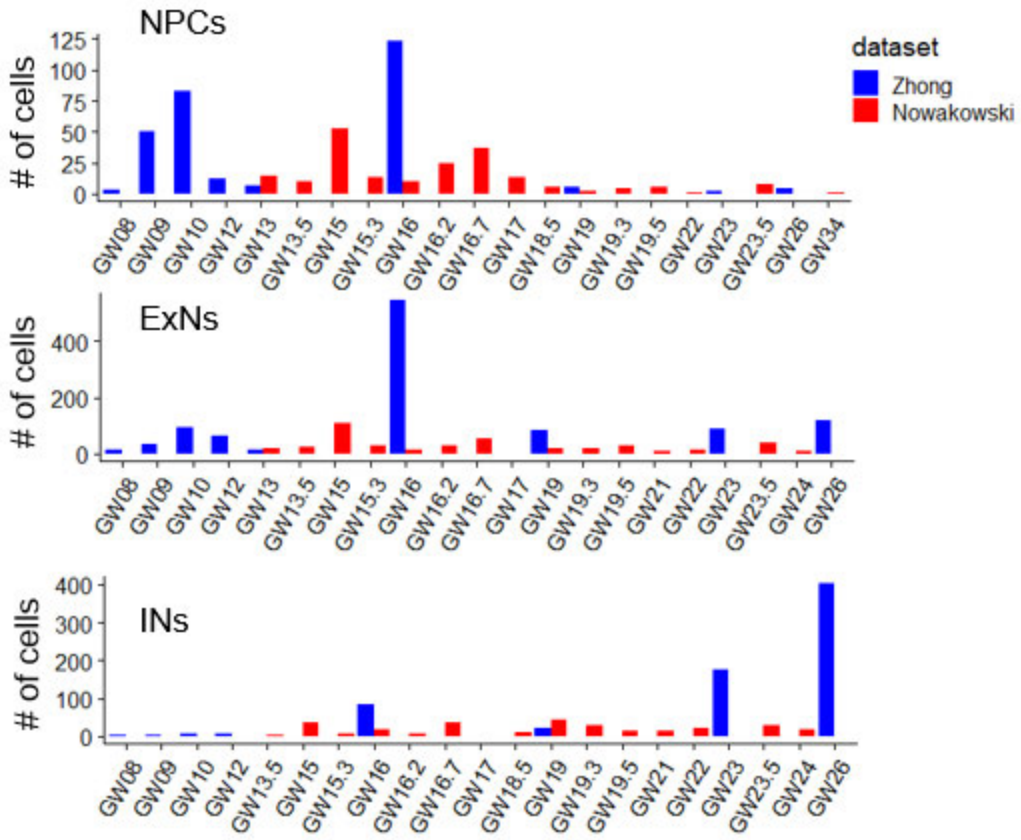


Figure 48: Bar plot showing the distribution of cell sampling in each cardinal cell class.

Blue, Zhong's dataset; Red, Nowakowski dataset.

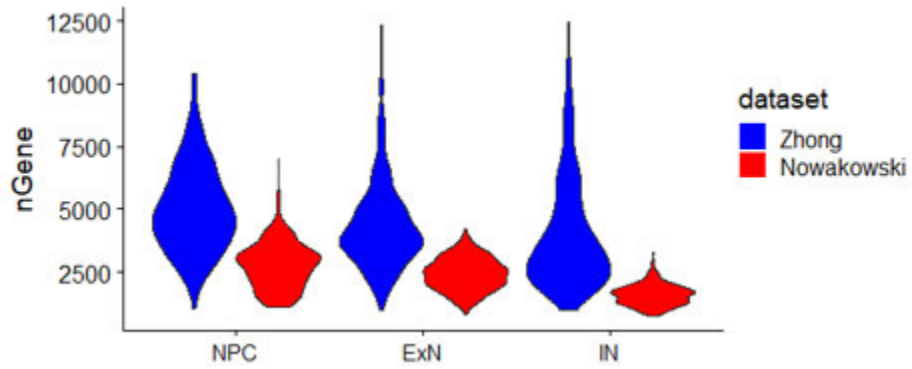


Figure 49: Violin plot indicating the difference of the number of genes detected per cell between two datasets.

Chapter 5: Enriched expression of genes associated with autism spectrum disorders in developing mouse interneurons

5.1 Introduction

Human developing PFC expansion likely contributed to the remarkable cognitive abilities of humans. In Chapter 3, we found that many of the ASD risk genes were differentially expressed in major/sub cell types that belong to Progenitor, Interneuron, Astrocyte and Microglia. Bioinformatics analysis revealed gene expression patterns at the single cell level that suggest some cells, most strikingly certain interneurons, may be disproportionally vulnerable to a large number of ASD causing mutations.

So, the interneurons were thought to primarily reflect the enrichment of ASD risk transcripts during human cortical development, and these ASD-affected genes may play roles in E/I balance, inhibitory neurogenesis and neuron-glia signalling. Due to the impossibility of wet lab experiments in embryonic human tissues, we asked if the two human clusters we identified in Chpater 3 (IN5 and IN8) matching any mouse interneuron clusters and if the matched mouse clusters also display enriched expression of ASD risk genes. Here, we searched for such differences by comparing embryonic cortical interneurons from human and mouse using single cell transcriptomics.

As described in Chapter 1, during embryonic brain development of both human and mouse, cortical interneurons are generated outside the cortex. To study cell diversity in the germinal regions of cortical interneurons, Da et al. dissected tissue from three regions in the mouse subpallium, including the dorsal and ventral medial ganglionic eminence (dMGE and vMGE, respectively) and the caudal ganglionic eminence (CGE) (Mi *et al.*, 2018). We hypothesised that there might be spatially and temporally distinct progenitor and precursor cell

populations in the embryonic GE regions, and each cohort of them were committed to generate certain cortical interneuron lineages during embryonic neurogenesis. So basically, we would like to find out the expression pattern of ASD risk genes among the interneuron progenitor and immature interneuron at GE regions, and the cortical interneuron diversity in the embryonic mouse.

5.2 Aim of this chapter

Since the interneurons were migrated from embryonic GE regions to cortex, firstly, we tried to illustrate the expression pattern of these genes across the interneurons in the medial and caudal ganglionic eminences. Then we looked at the diversity of cortical interneurons in embryonic mouse. We also tried to assign the embryonic interneurons to the adult interneuron lineages to reveal the embryonic interneuron linages. Finally, we tested if the human interneuron cell type that we identified in Chapter 3 matched any mouse interneuron clusters and if the matched mouse clusters display enriched expression of ASD risk genes.

5.3 Materials and methods

We used two scRNA-seq datasets in this Chapter. The first one was mouse cells collected from E12.5 and E14.5 MGE (dorsal and ventral) and CGE regions (Mi *et al.*, 2018). 2,003 single cells were isolated and subjected to cDNA synthesis and RNA-seq using a Fluidigm C1 system. The second one was Lhx6+ cortical interneurons collected by fluorescence-activated cell sorting (FACS) from E18.5 mouse cortex. 2,432 single cells were sequenced by Drop-seq system. In this chapter, we used the authors' original classification result of interneuron cell types. The transcript counts of cells in both datasets were downloaded from original publication, and normalized to counts per

million (CPM), where CPM is the transcript count of each gene divided by the sum of transcript counts of that cell, multiplied by one million.

All methods of scRNA-seq analysis were the same as we described in Chapter 3.

5.4 Results

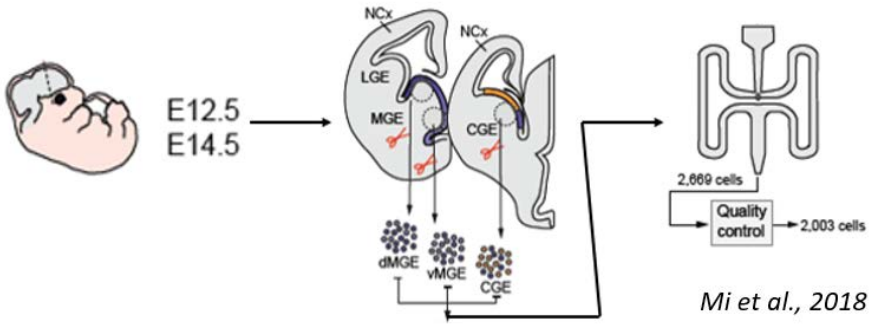
5.4.1 Cellular heterogeneity of interneurons in the developing mouse ganglionic eminences

It has been well established that cortical interneurons are developmentally derived from progenitor domains in embryonic subpallium areas, including MGE and CGE, that give rise to different types of cortical interneurons (Xu, Tam and Anderson, 2008; Gelman and Marín, 2010; Melzer *et al.*, 2017). To determine if ASD risk genes are expressed broadly or specifically in cortical interneuron precursor cells during early brain development, we carried out a single-cell RNA-seq survey of interneuron precursor cells (progenitor cells and newborn interneurons), in a collaboration with Oscar Marin's lab at the King's College London, to explore the expression pattern of both monogenic and *16p11.2* ASD risk genes at the single-cell level in mouse embryonic subpallium.

In this study, Mi et al. manually dissected ganglionic eminence cells from E12.5 and E14.5 mouse embryos (Figure 50A). The developmental time point we chose in this study corresponded to the peak of cortical interneuron neurogenesis in mouse embryo over the interval E12.5-E14.5. Mi et al. applied Fluidigm C1 system to generate single-cell gene expression profiles of over 2000 single cells collected from dorsal MGE (dMGE), ventral MGE (vMGE) and CGE regions that are thought to give rise to different types of cortical interneurons according to previous studies (Inan, Welagen and Anderson,

2012). Hierarchical clustering of all collected cells based on their gene expression profiles confirmed the presence of both mitotic progenitor cells and postmitotic interneurons in this scRNA-seq dataset (Figure 50B).

(A)



(B)

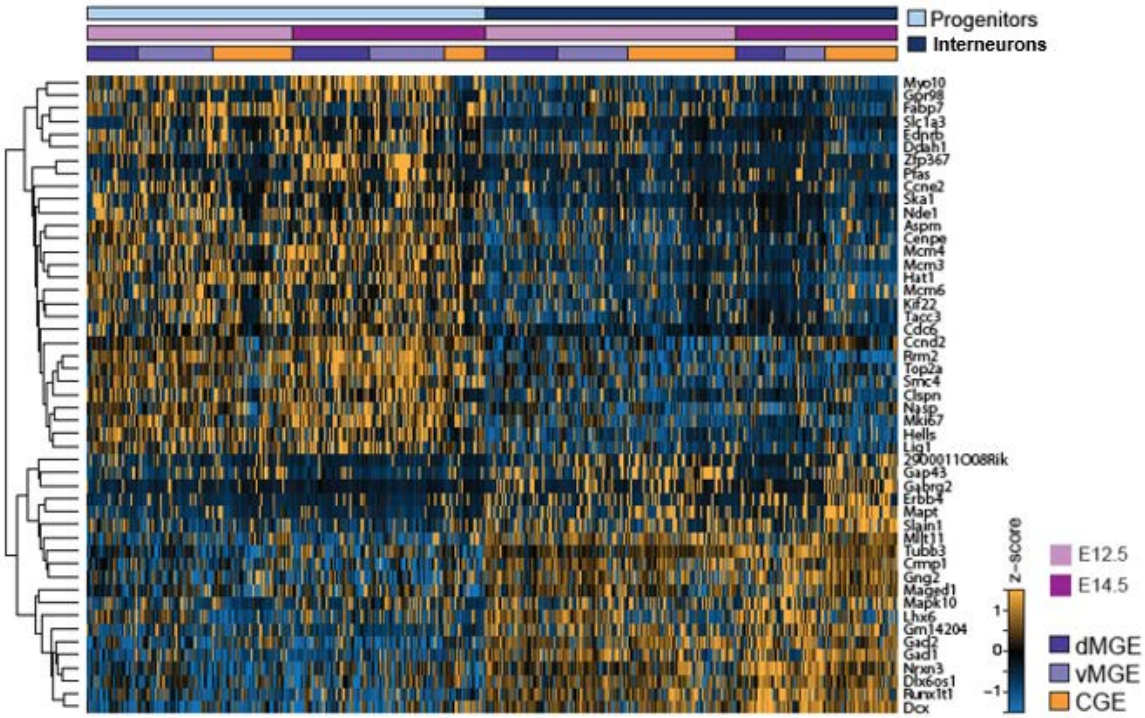


Figure 50: Major sources of transcriptional heterogeneity among single cells from mouse MGE and CGE.

(A) Schematic illustrating sample collection, sequencing and scRNA-seq analysis workflow. Single cells from E12.5 and E14.5 dMGE, vMGE and CGE were isolated and subjected to cDNA synthesis and RNA-seq using a Fluidigm C1 system. (B) The heatmap illustrating average expression of genes selected that best represent progenitor or neuronal identity (Mi et al., 2018). Coloured bars above heatmap identify cell identity, stage and region of origin.

We next turned our attention to explore the expression of ASD risk genes in progenitor cells. Firstly, we applied an unsupervised approach to classify progenitor cells at two developmental stages. This analysis identified 13 and 11 transcriptionally distinct progenitor cell clusters at E12.5 and E14.5 respectively (Figure 51). The clustering analysis of progenitor cells confirmed that embryonic subpallium contains a highly dynamic pools of progenitor cells at different developmental stages and progenitor domains. Differential gene analysis further illustrated the transcriptomic signatures of progenitor cell clusters, which defines their ventricular zone (VZ) radial glial cell and subventricular zone (SVZ) intermediate progenitor identities along with their region of origin (dMGE, vMGE and CGE) (Figure 52). We then systematically examined the expression of ASD risk genes in progenitor cell clusters at E12.5 and E14.5 respectively (Figures 52 to 57).

5.4.1.1 ASD gene expression in IN progenitors

A comprehensive expression profile of both monogenic and *16p11.2* ASD risk genes among progenitor cell clusters at two ages were illustrated by violin plots (Figure 53, 54, 56 and 57). We found that the majority of ASD risk genes we examined are broadly expressed in all progenitor clusters with only a few exceptions. To better evaluate the cell type specificity of ASD risk genes in progenitor clusters, we conducted differential gene expression analysis across progenitor clusters at both developmental stages. Surprisingly, no significant enrichments of ASD risk genes were found in any of E12.5 progenitor cell clusters, while 5 ASD risk genes are enriched in different E14.5 progenitor cell clusters (Figure 55). Among 5 ASD risk genes with cell-type specific enrichment pattern, *Med13l, Dyrk1a, Bcl11a* and *Ypel3* are enriched in SVZ intermediate progenitor clusters (P5, P6 and P11), while only *Aldoa* gene is enriched in VZ radial glial cell cluster (P1) (Figure 55 A).

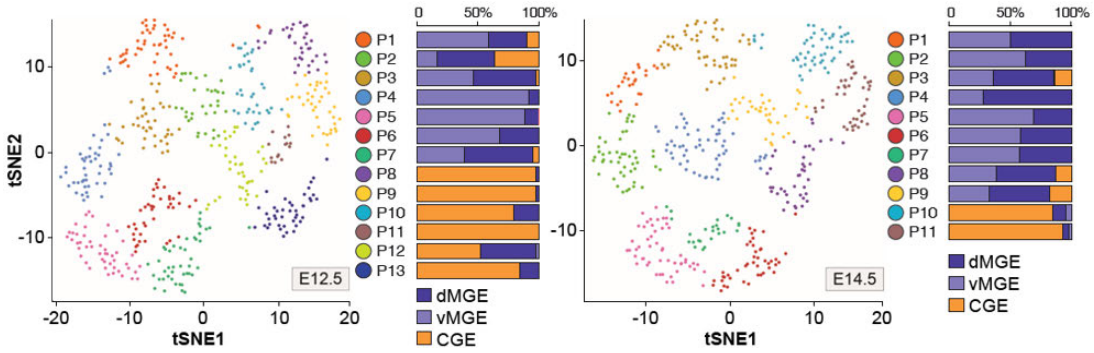


Figure 51: Visualization of progenitor cell diversity at E12.5 (left) and E14.5 (right) by t-SNE.

Histograms illustrate the relative contribution of dMGE, vMGE and CGE cells to each progenitor cluster.

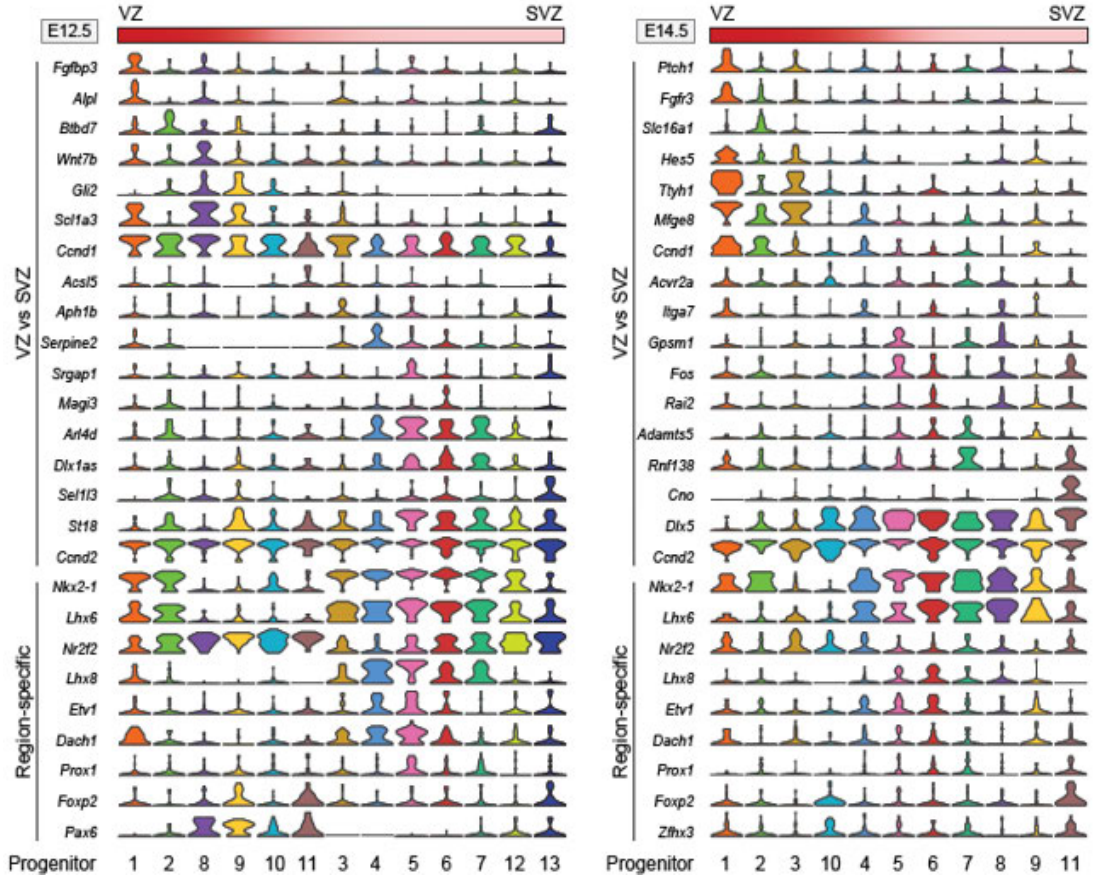


Figure 52: Violin plots depicting the expression of marker genes that distinguish VZ/SVZ identities and patterning information in progenitor clusters at E12.5 (left) and E14.5 (right).

Known and novel markers enriched in individual or multiple clusters were selected. *Hes5* and *Slc1a3* are known markers of radial glial cells across multiple regions of the developing telencephalon. *Ccnd1* is also enriched in VZ cells. *Gli2*, *Fgfbp3*, *Snp23*, *Tyh1*, *Fgfr3* and *Mfge8* are newly identified markers for VZ progenitor cell clusters. *Dlx6* is known to mark intermediate progenitor cells in the SVZ and newborn interneurons, while *St18* is a novel marker of progenitor cell clusters in the SVZ. *Nkx2-1* and *Lhx6* are markers of MGE identity. *Etv1* and *Lhx8* are enriched in vMGE progenitor cells, while *Nr2f2* marks progenitor cells in dMGE and CGE. *Dach1* is a newly identified vMGE progenitor cell marker. *Pax6* is a well-known marker for CGE progenitors. *Prox1*, *Foxp2* and *Zfhx3* are markers of CGE progenitor cells.

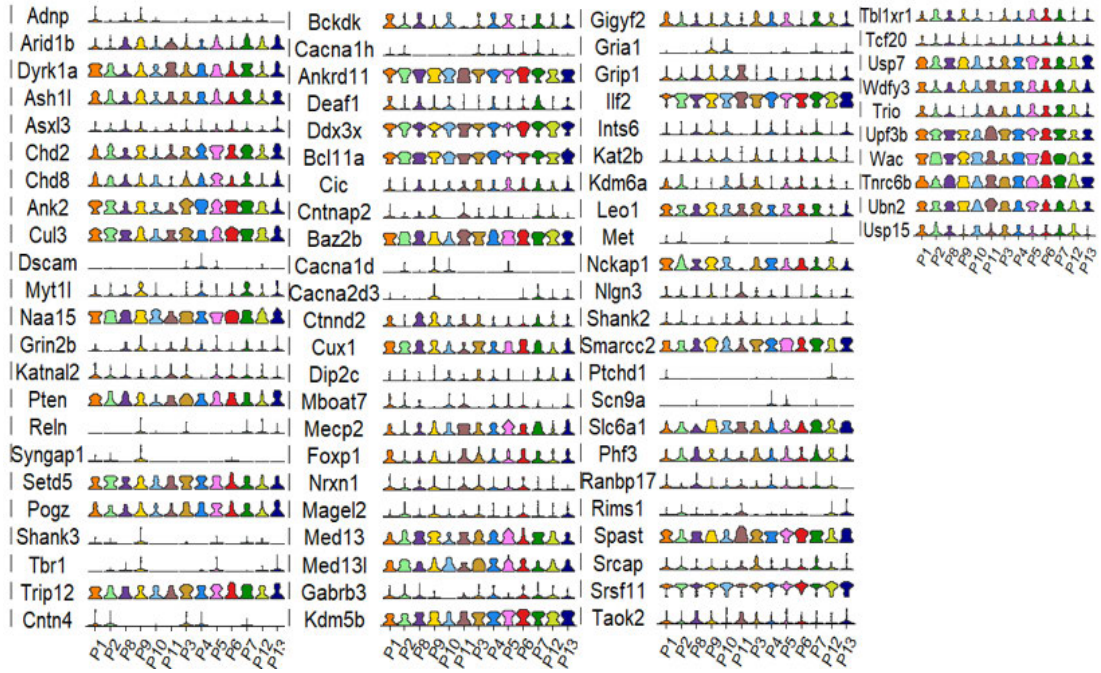


Figure 53: Violin plot illustrating expression pattern of monogenic ASD risk genes across thirteen clusters of E12.5 mouse progenitors.

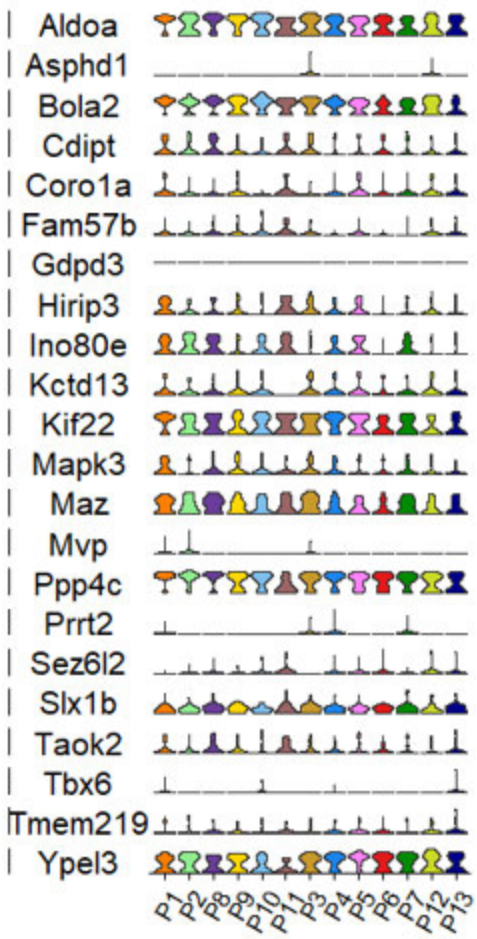


Figure 54: Violin plot illustrating expression pattern of ASD risk genes on *16p11.2* locus across thirteen progenitor clusters of E12.5 mouse progenitors.

A

	avg_logFC	p_val_adj	cluster
Med13l	0.884523282	0.002355561	P5
Dyrk1a	0.695440035	0.018372624	P5
Bcl11a	1.104891689	0.012287231	P6
Aldoa	1.560795221	0.035469042	P1
Ypel3	1.369478045	0.000188274	P11

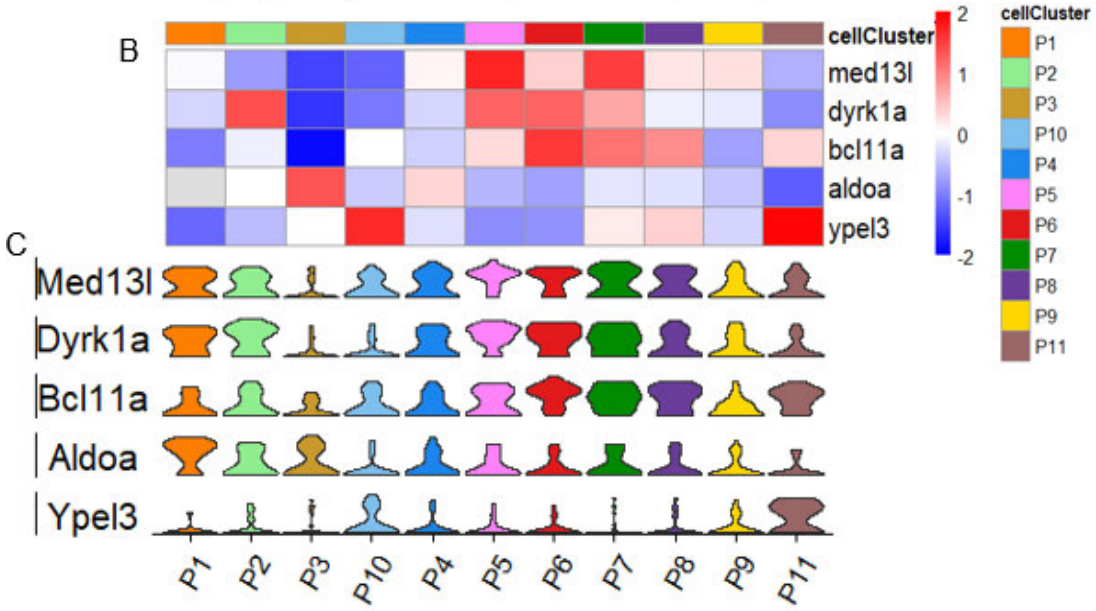


Figure 55: The differential expressed ASD risk genes across E14.5 progenitor clusters.

(A) Table illustrating the enrichment of significantly differential expressed ASD risk genes across progenitor clusters. Green box: monogenic ASD risk genes; yellow box: CNV genes on *16p11.2* locus (Wilcox test, adjust p value < 0.05, log (fold change) > 0.3). (B and C) Heatmap (B) and violin plot (C) illustrating the expression pattern of significantly differential expressed ASD risk genes across cell clusters.

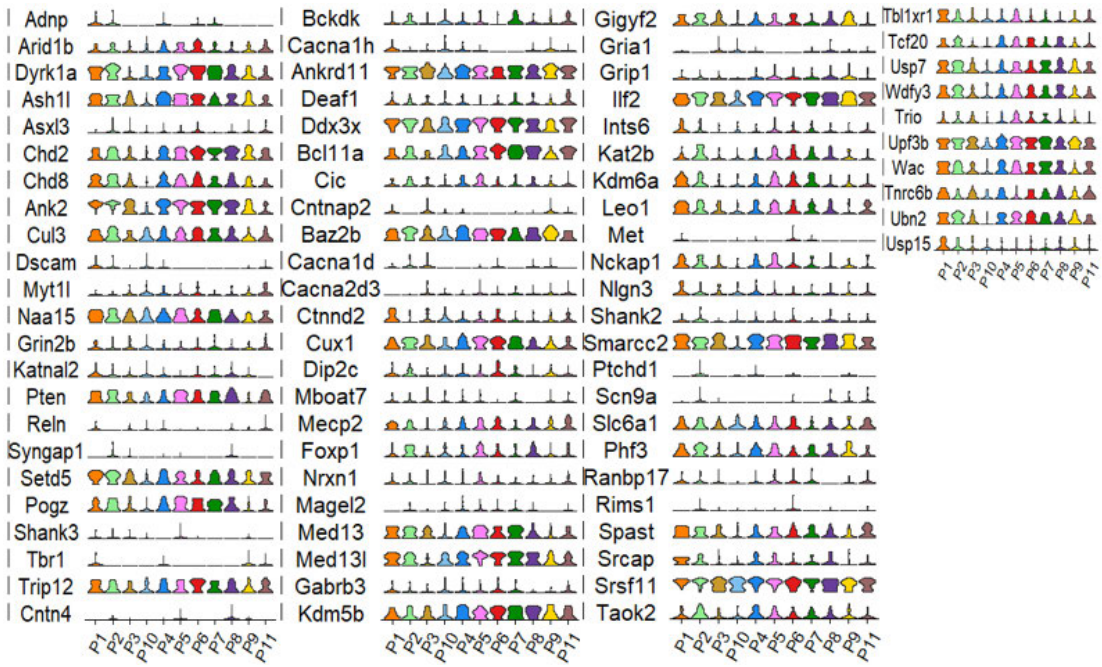


Figure 56: Violin plot illustrating expression pattern of monogenic ASD risk genes across eleven clusters of E14.5 mouse progenitors.

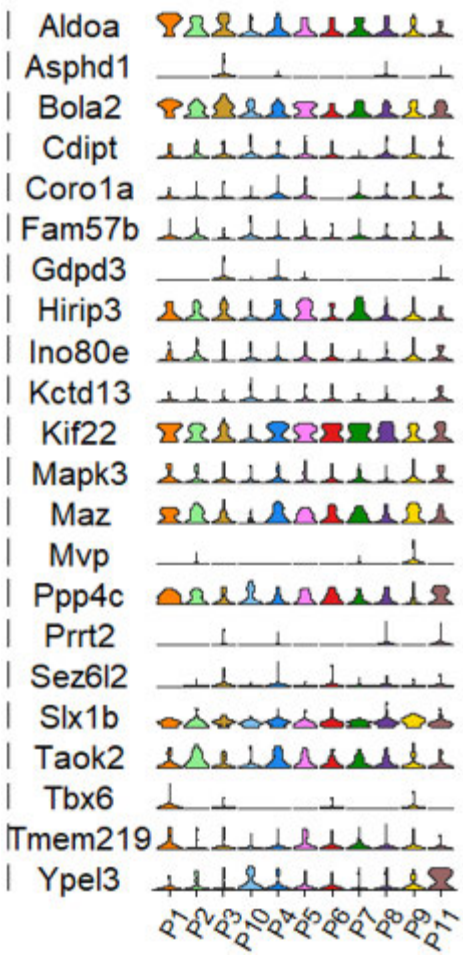


Figure 57: Violin plot illustrating expression pattern of ASD risk genes on *16p11.2* locus across eleven progenitor clusters of E14.5 mouse progenitors.

5.4.1.2 ASD gene expression in newborn INs

We then sought to evaluate the expression pattern of ASD risk genes in newborn interneurons in mouse embryonic subpallium. We first applied unsupervised clustering analysis on newborn interneurons at both E12.5 and E14.5. Unbiased clustering analysis identified 13 clusters of newborn interneurons with distinctive gene expression profiles, as well as specific temporal identities (Figure 58). This analysis revealed that temporal identity

segregates more clearly among interneuron clusters (figure 58 B and C), indicating that interneurons become more transcriptionally heterogeneous over embryonic brain development. It has been well established that the MGE and CGE generate different groups of cortical interneurons (Xu, Tam and Anderson, 2008; Gelman and Marín, 2010; Melzer *et al.*, 2017). Most Pv+ and Sst+ interneurons are born in the MGE, whereas the CGE is the origin of Vip+ interneurons and neurogliaform (Ndnf+) cells (Kelsom and Lu, 2013). To better annotate 13 interneuron clusters with lineage identities, we hierarchically organize them based on the expression of a set of region- and interneuron cell type- specific marker genes (Figure 59). This analysis revealed a prominent segregation of MGE-derived interneuron clusters from CGE-derived interneuron clusters, but less clear segregation in terms of their cell type identities, which is largely due to the lack of the expression of mature interneuron cell type markers in newborn interneurons (Figure 59).

We then systematically examined the expression of ASD risk genes in 13 interneuron clusters. A comprehensive expression profile of both monogenic and *16p11.2* ASD risk genes among interneuron clusters were illustrated by violin plots (Figure 61 and 62). Differential gene expression analysis further confirmed the significant enrichment of 14 ASD risk genes in distinct interneuron clusters (IN1, IN3, IN5, IN8, IN12 and IN13) (Figure 60). Notably 6 out of 14 differentially expressed ASD risk genes are enriched in IN5, suggesting the hypothesis that these cells may represent a convergent target for multiple ASD causing mutations acting during embryonic brain development.

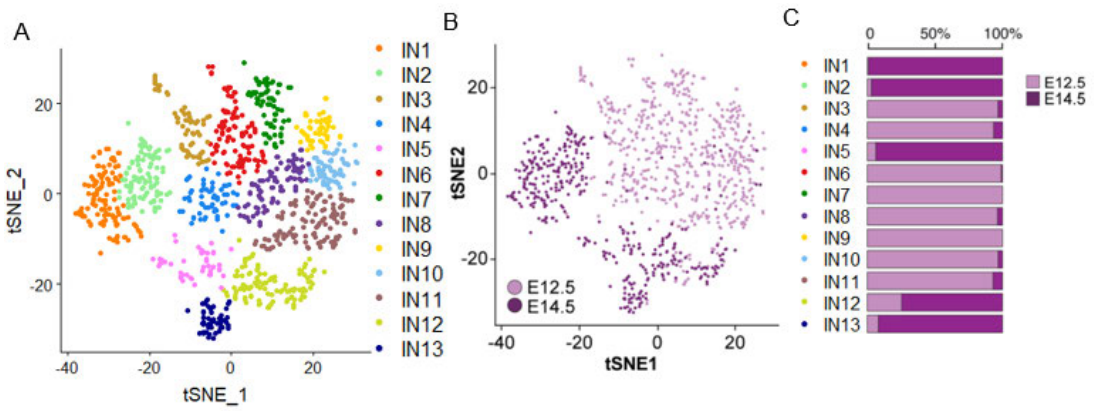


Figure 58: Emergence of cortical interneuron diversity in the ganglionic eminences.

(A) *t*-SNE plot depicting neuronal clusters following unsupervised clustering. (B) E12.5 and E14.5 interneurons are depicted in the same t-SNE space. (C) Histograms illustrate the relative contribution E12.5 and E14.5 cells to each interneuron cluster.

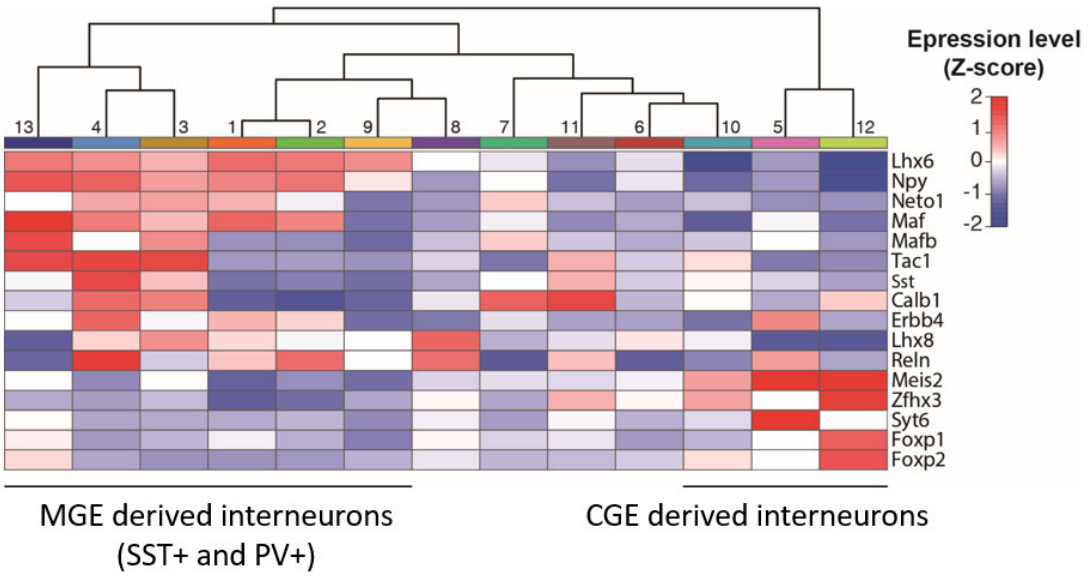


Figure 59: The heatmap illustrates average expression of known interneuron lineage associated genes in the newly identified neuronal clusters.

Lhx6, Npy, Neto1, Maf, Mafb and *Calb1* are enriched in MGE-derived interneuron lineages (SST+ and PV+ interneurons), while *Meis2, Zfhx3, Syt6, Foxp1* and *Foxp2* are enriched in CGE-derived interneuron lineages (VIP+ and NDNF+).

A

	avg_logFC	p_val_adj	cluster
Gria1	2.499256441	5.08E-12	IN13
Slc6a1	0.717192062	0.012221392	IN13
Ddx3x	0.658751834	0.02254048	IN3
Med13l	1.128347632	1.92E-05	IN1
Chd8	1.313715856	0.019655784	IN1
Ddx3x	0.544661125	0.012762197	IN8
Grin2b	2.151787707	1.06E-14	IN5
Foxp1	1.861570617	1.17E-12	IN5
Gria1	0.868526032	3.28E-08	IN5
Myt1l	0.817999599	0.001076631	IN5
Deaf1	0.992218628	0.037050392	IN5
Ash1l	0.326951328	0.04706046	IN5
Foxp1	1.423191051	2.03E-06	IN12
Sez6l2	1.148722645	8.07E-06	IN13

B

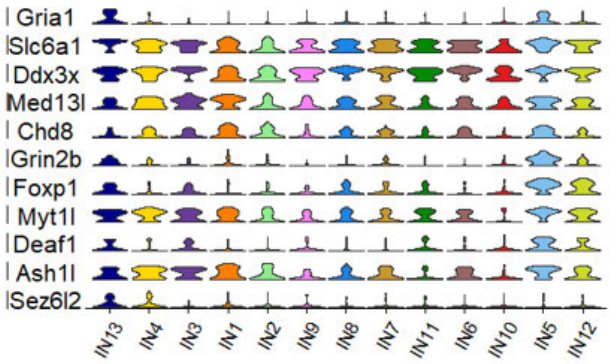


Figure 60: The differential expressed ASD risk genes across interneuron clusters.

(A) Table illustrating the enrichment of significantly differential expressed ASD risk genes across interneuron clusters. Green box: monogenic ASD risk genes; yellow box: CNV genes on 16p11.2 locus (Wilcox test, adjust p value < 0.05, log (fold change) > 0.3). (B) Violin plot illustrating the expression pattern of significantly differential expressed ASD risk genes across interneuron clusters.

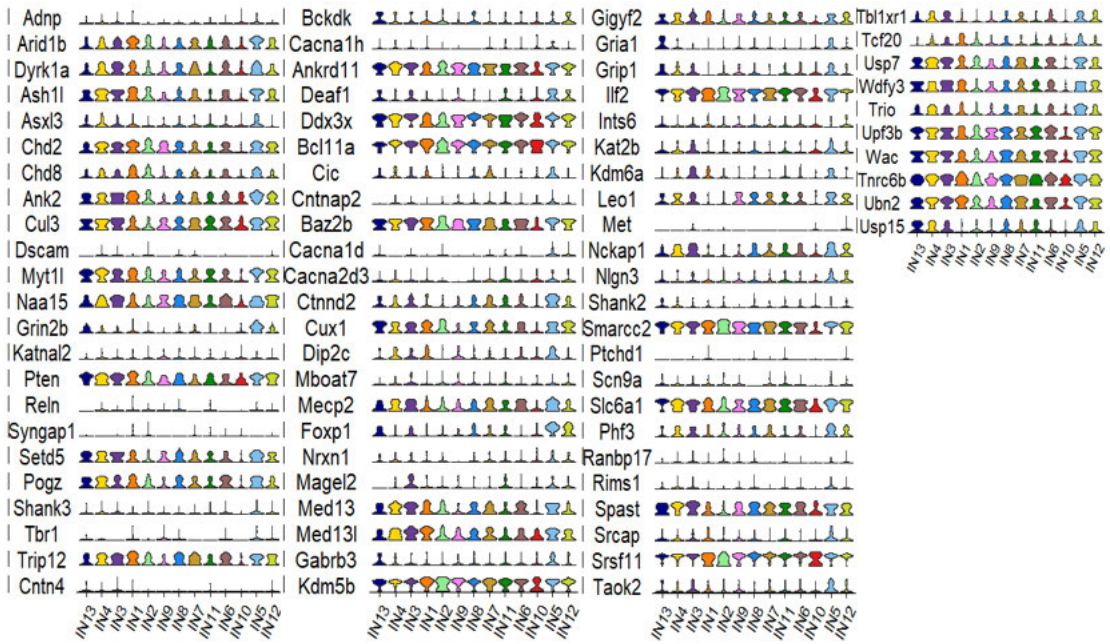


Figure 61: Violin plot illustrating expression pattern of monogenic ASD risk genes across twelve interneuron clusters.

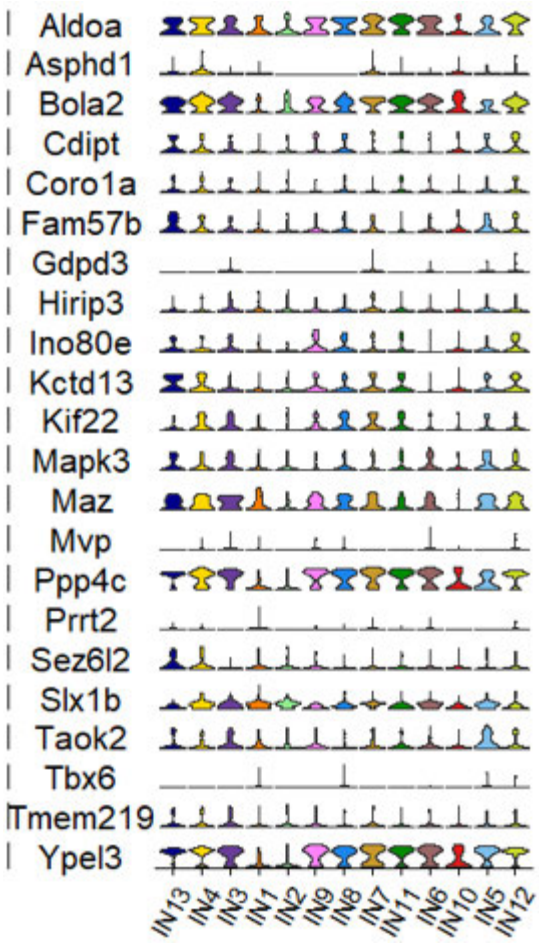
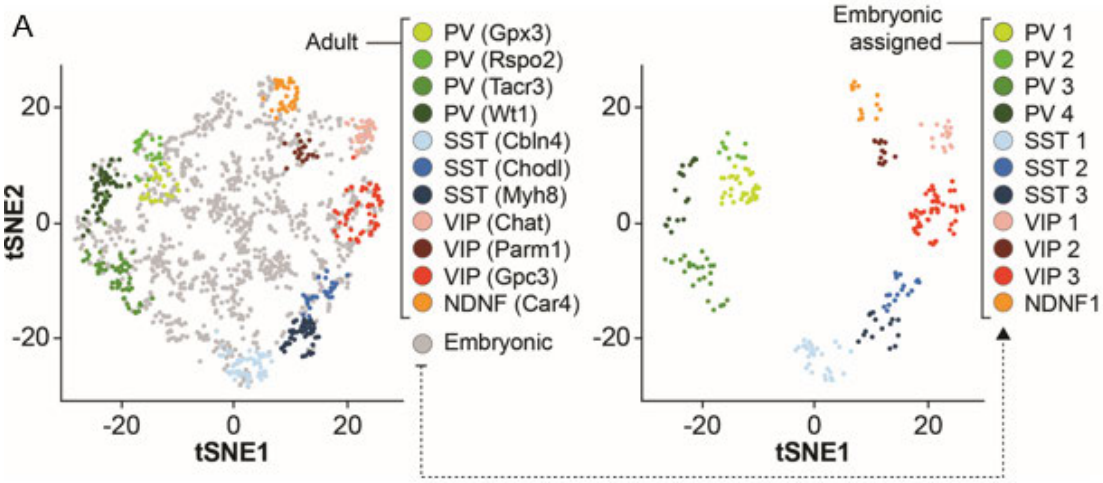


Figure 62: Violin plot illustrating expression pattern of ASD risk genes on *16p11.2* locus across twelve interneuron clusters.

The lack of expression of mature interneuron cell type marks in embryonic newborn interneurons makes it difficult to define their cell type identities. To overcome this issue, we used a publicly available scRNA-seq dataset of 761 mature cortical interneurons from adult mouse visual cortex as a reference dataset (Tasic *et al.*, 2018) and applied canonical correlation analysis (CCA) to assign embryonic newborn interneurons into 11 interneuron cell types according to their transcriptomic similarity to mature interneurons from the reference dataset (Figure 63A). The robustness of the cell type assignment analysis was assessed by the MetaNeighbor analysis (Figure 63B), which confirmed good matching between assigned embryonic interneuron classes

and adult cortical interneuron cell types based on high AUROC values (above 0.7). These assigned 11 embryonic interneuron classes correspond to anatomically and electrophysiologically defined 4 major classes of cortical interneurons, including MGE derived SST+ and PV+ interneuron, as well as CGE derived VIP+ basket and bipolar interneurons, and NDNF+ neurogliaform cells (Figure 64A). Analysis of the contribution of cell type identities to 11 interneuron clusters identified by unbiased clustering method showed that 4 major interneuron classes all contributed to every interneuron clusters (Figure 64B). This result suggests that interneuron clusters identified by unbiased clustering method do not represent a particular interneuron type but rather a cell-state that interneurons of multiple lineages (eg SST+, PV+, VIP+ and NDNF+) go through *en route* to reaching maturity. Finally, we examined the expression of both monogenic and *16p11.2* ASD risk genes in 4 major interneuron classes defined by the cell assignment analysis, which is illustrated by violin plots (Figure 65 and 66). We found that most but not all ASD risk genes are broadly expressed in multiple interneuron classes without obvious cell type specificity. This is further confirmed by the following differential gene expression analysis showing that no significant enrichment of ASD risk genes was found in 4 major interneuron classes.



B

Adult	CCA embryonic assigment	AUROC
PV (Gpx3)	PV1	0.79
PV (Rspo2)	PV2	0.91
PV (Tacr3)	PV3	0.98
PV (Wt1)	PV4	0.93
SST (Cbln4)	SST1	0.86
SST (Myh8)	SST2	0.95
SST (Chodl)	SST3	1.00
VIP (Chat)	VIP1	0.85
VIP (Parm1)	VIP2	1.00
VIP (Gpc3)	VIP3	0.93
NDNF (Car4)	NDNF1	0.76

Figure 63: Integration of embryonic neurons and adult cortical interneurons in t-SNE space.

(A) Embryonic neurons (right) assigned to specific interneuron lineages (left) are depicted in the same t-SNE space. Unassigned embryonic neurons are omitted. (B) The table shows mean AUROC scores between assigned embryonic interneuron classes and adult cortical interneuron cell types. All mean AUROC scores are above 0.7 that typically suggest the good correlation.

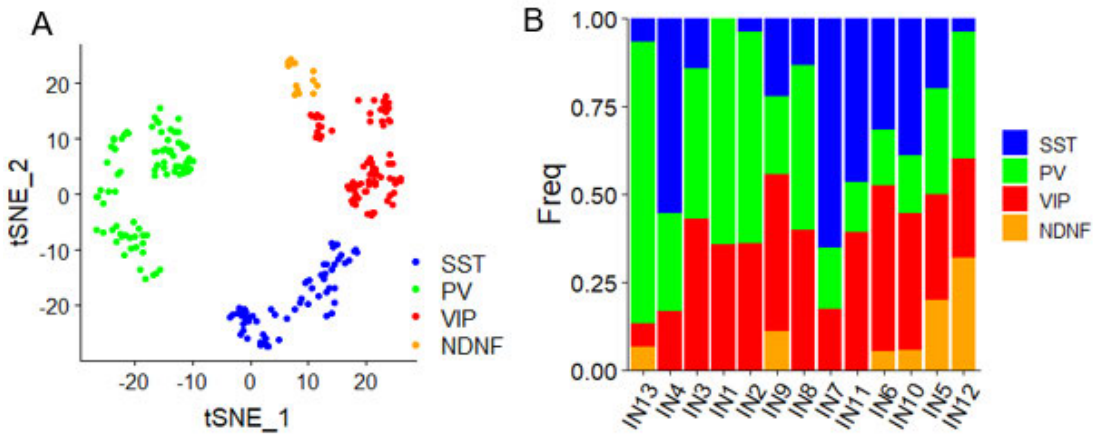


Figure 64: Integration of embryonic neurons and adult cortical interneurons in t-SNE space.

(A) Embryonic neurons (right) assigned to specific interneuron lineages (left) are depicted in the same t-SNE space. (B) Histograms illustrating the relative contribution of interneuron lineages to each interneuron cluster.

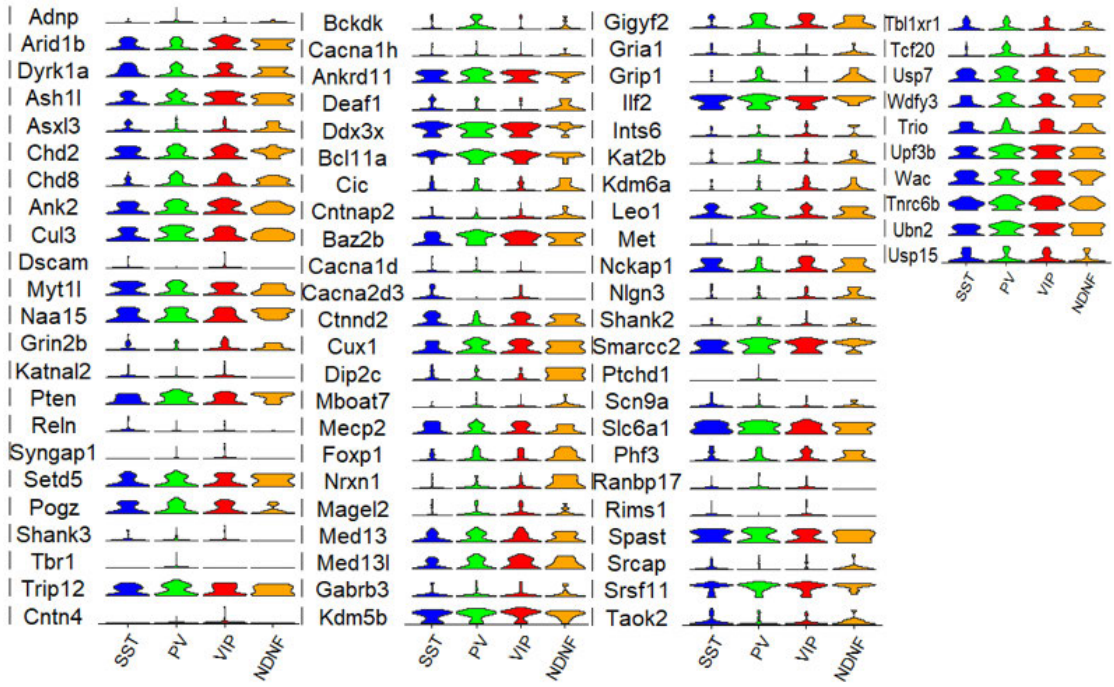


Figure 65: Violin plot illustrating expression pattern of monogenic ASD risk genes across four major interneuron classes.

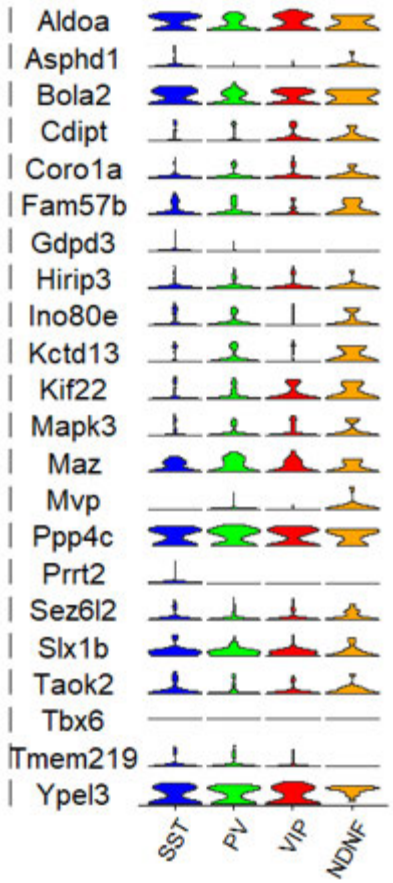


Figure 66: Violin plot illustrating expression pattern of ASD risk genes on *16p11.2* locus across four major interneuron classes.

5.4.2 Cellular heterogeneity of interneurons in the developing mouse cortex

To further examine the expression patterns of ASD risk genes in cortical interneurons of the developing mouse cortex, we employed a publicly available scRNA-seq dataset which comprised of interneurons isolated from E18.5 mouse cortex (Figure 67A). Seven non-overlapping cell types of cortical interneurons (Sst, Nos1, Th, Pvalb, Vip, Id2 and Igfbp6) were identified in this

dataset by cell type alignments across scRNA-seq datasets of embryonic and adult mouse cortex (Figure 67B) (Mayer *et al.*, 2018).

We carefully examined the expression pattern of both monogenic and *16p11.2* ASD risk genes across 7 major interneuron cell types and found out that the majority of ASD risk genes are broadly expressed in multiple interneuron cell types with only a few exceptions (Figure 68 and 69). Differential gene expression analysis further confirmed that 7 ASD risk genes are significant enriched in either Sst+ or Pvalb+ interneuron types respectively (Figure 70). In detail, *Reln, Bcl11a, Cacna2d3, Nrxn1* and *Ctnnd2* genes were significantly enriched in Sst+ interneurons. *Cux1* gene was significantly enriched in Pvalb + interneurons. And *Gria1* gene was significantly enriched in both Sst+ and Pvalb+ interneurons.

A

cellType	# of cells
Sst	718
Nos1	7
Th	17
Pvalb	524
Vip	608
Id2	536
Igfbp6	22

B

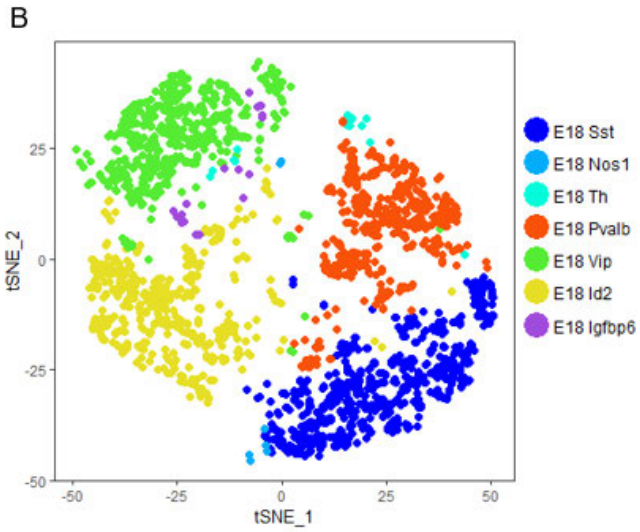


Figure 67: Cellular heterogeneity of interneurons in the developing mouse cortex.

(A) Table summarizing the number of cells in each cell type. (B) *t*-SNE plot illustrating the cell types of cortical interneurons (Sst, Nos1, Th, Pvalb, Vip, Id2 and Igfbp6) in this dataset.

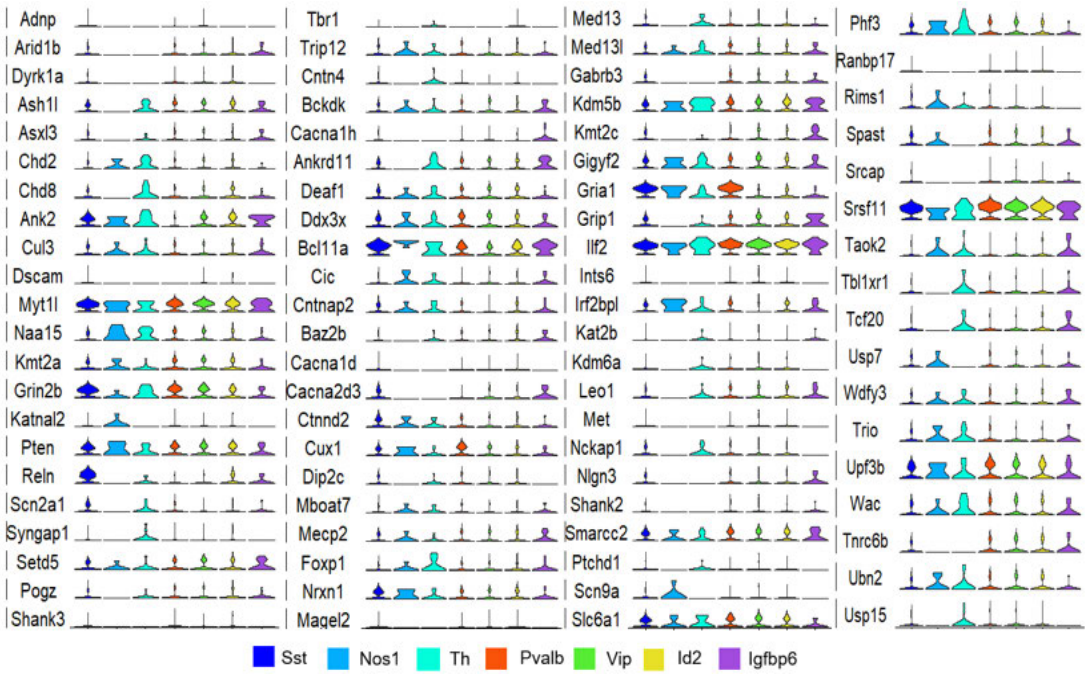


Figure 68: Violin plot illustrating expression pattern of monogenic ASD risk genes across seven interneuron cell types.

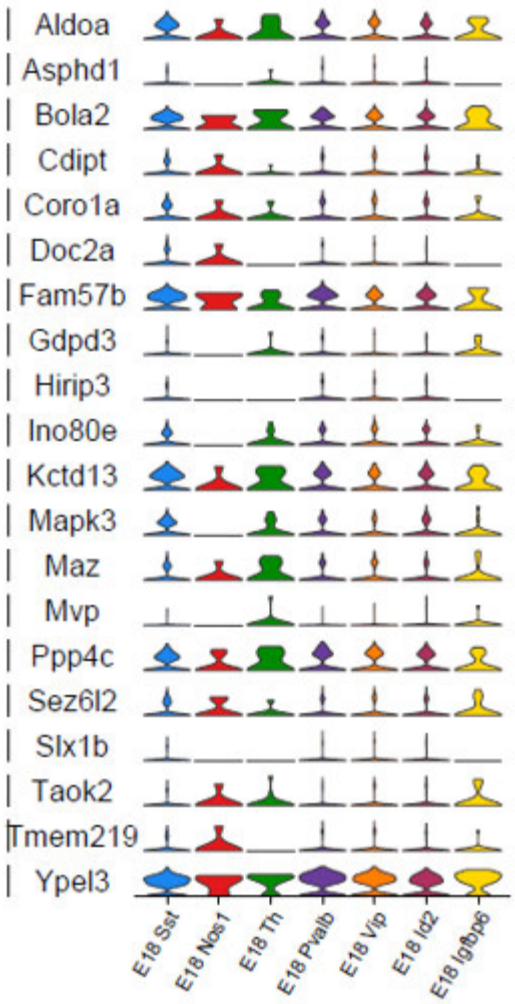


Figure 69: Violin plot illustrating expression pattern of ASD risk genes on *16p11.2* locus across seven interneruon cell types.

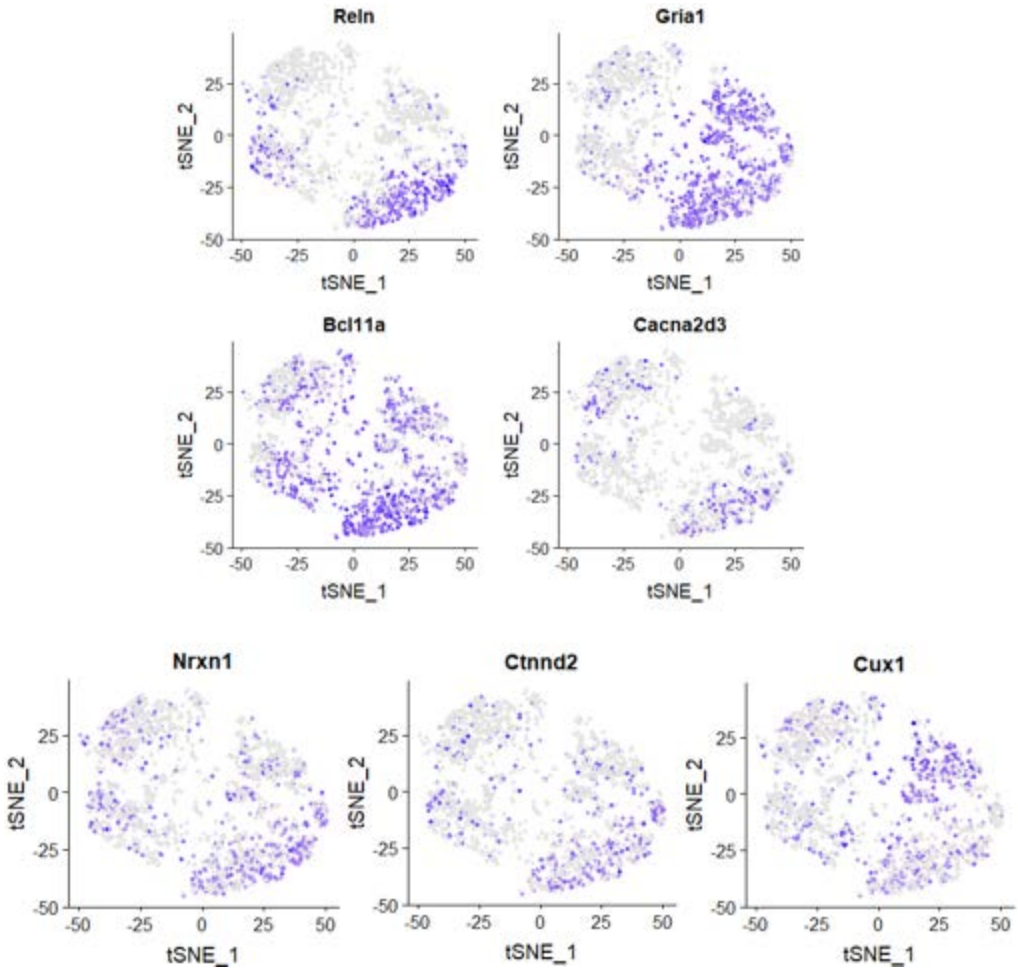


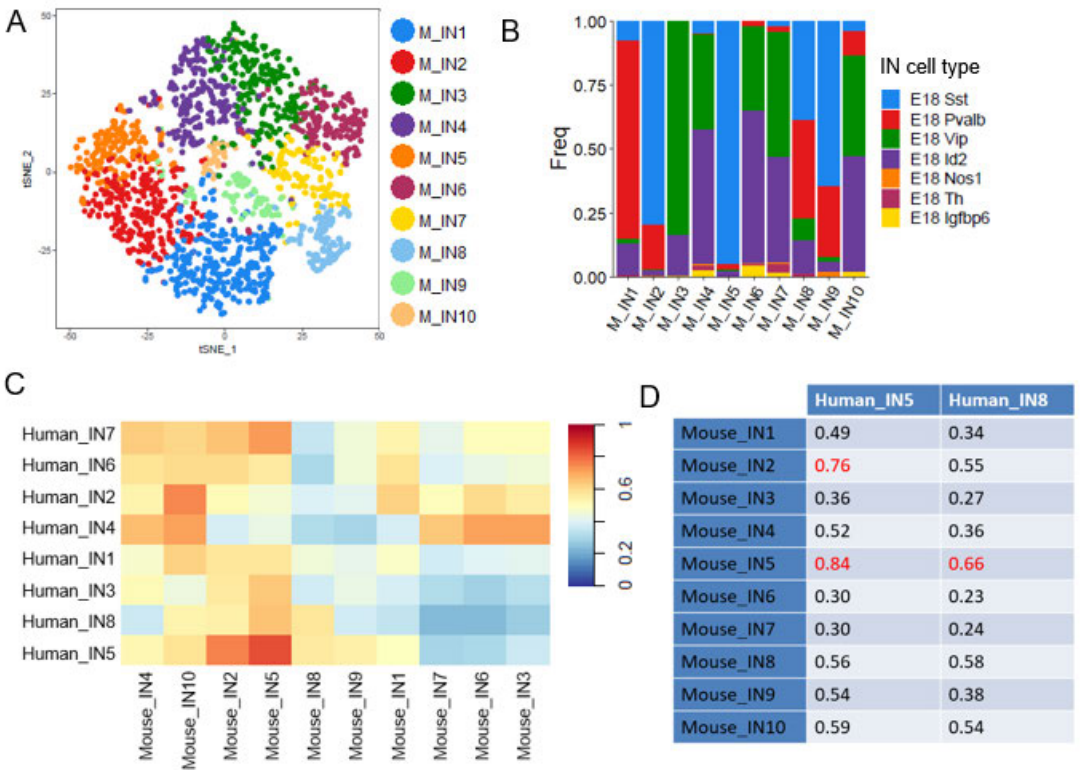
Figure 70: Gradient plot showing the expression pattern of the significantly differentially expressed ASD risk genes.

5.4.2.1 Identfying mouse developing IN correlates of human developing INs

We performed unsupervised clustering on E18.5 cortical interneurons and classified them into ten transcriptionally distinct cell clusters (Figure 71A). We quantified the proportion of seven cardinal interneuron cells types in each of interneuron clusters and found that all of the interneuron clusters contains multiple interneuron cell types (Figure 71B), indicating that some aspects of the interneuron diversity at the embryonic stage might not link to their cell type identities. We also conducted MetaNeighbor analysis to examine the conservation of interneuron diversity identified in E18.5 mouse cortex and human fetal cortex (Figure 71C and D). The degree of conservation was determined based on shared gene expression patterns between species illustrated by the heatmap of AUROC values (Figure 71C). This analysis revealed homologous cell clusters of human interneuron clusters IN5 and IN8 in E18.5 mouse cortex and highlighted two matching mouse interneuron clusters M_IN2 and M_IN5 according to high AUROC values (above 0.6) (Figure 71D).

We showed that human interneuron clusters IN5 and IN8 exhibit the enrichment of ASD risk gene expression in Chapter 3. Now we asked if the two human IN5 and IN8 clusters matching any mouse interneuron clusters and if so whether the matched mouse clusters also display enriched expression of ASD risk genes. We examined the expression pattern of both monogenic and 16p11.2 ASD risk genes among mouse interneuron clusters (Figure 72 and 74). The following differential gene expression analysis confirmed a significant enrichment of ASD risk gene expression in mouse interneuron clusters M_IN2 and M_IN5 (Figure 72A). Interestingly, we noticed that a number of differentially expressed ASD risk genes are also enriched in interneuron cluster M_IN8 and M_IN9, which may suggest a divergence of ASD risk gene expression pattern between mouse and human cortex. We also carried out GO term enrichment analysis to identify GO terms that best discriminate

mouse interneurons clusters (Figure 72B). Interestingly, many of these GO terms that also discriminated interneuron clusters in the human fetal cortex mouse interneurons were the same as those that discriminated interneuron clusters in human fetal cortex (as mentioned in Chapter 3), which further highlighted the conservation of interneuron diversity between two species.



	Human_IN5	Human_IN8
Mouse_IN1	0.49	0.34
Mouse_IN2	0.76	0.55
Mouse_IN3	0.36	0.27
Mouse_IN4	0.52	0.36
Mouse_IN5	0.84	0.66
Mouse_IN6	0.30	0.23
Mouse_IN7	0.30	0.24
Mouse_IN8	0.56	0.58
Mouse_IN9	0.54	0.38
Mouse_IN10	0.59	0.54

Figure 71: Unsupervised clustering on E18.5 cortical interneurons and comparison between human and mouse cortical interneuron clusters.

(A) *t*-SNE plot showing the transcriptionally distinct cell clusters in this dataset. (B) Bar plot depicting the percentage of interneuron cell types in each mouse interneuron clusters. (C) Heatmap of AUROC values indicating the transcriptionally similarity between human IN clusters (see Chapter 3) and mouse clusters. In the colour bar, blue corresponds to low AUROC scores; red correspond to high low AUROC scores. (D) AUROC scores in the table indicating the correlation between the ten mouse interneuron clusters and two human interneuron clusters. An AUROC score of 0.6 or above suggests a high correlation, and these scores are coloured as red.

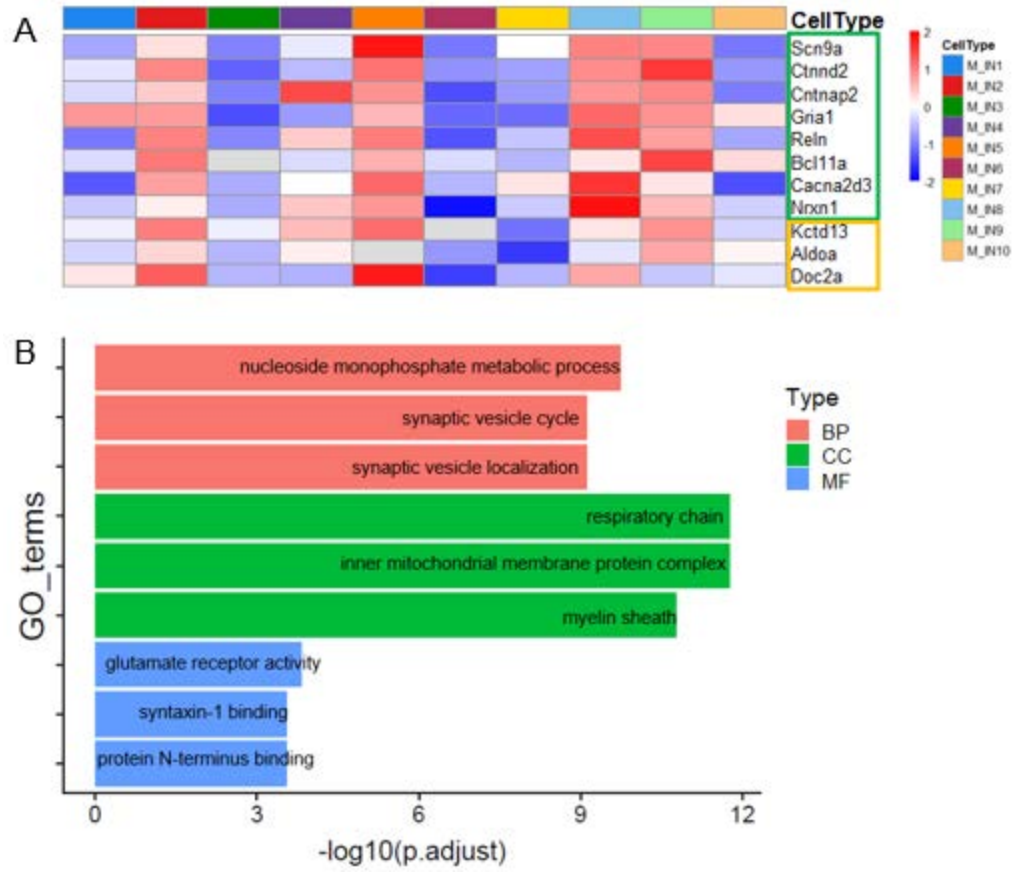


Figure 72: The diversity of mouse interneuron clusters.

(A) Heatmap illustrating the expression pattern of significantly differential expressed ASD risk genes across cell clusters. Green box: monogenic ASD risk genes; yellow box: CNV genes on *16p11.2* locus (Wilcox test, adjust p value < 0.05, log (fold change) > 0.3). (B) Top significant GO terms associated with the enriched DEGs across mouse interneuron clusters. Clusters belong to the different types of terms are color-coded accordingly. BP, biological processes; CC, cellular components; MF, molecular functions.

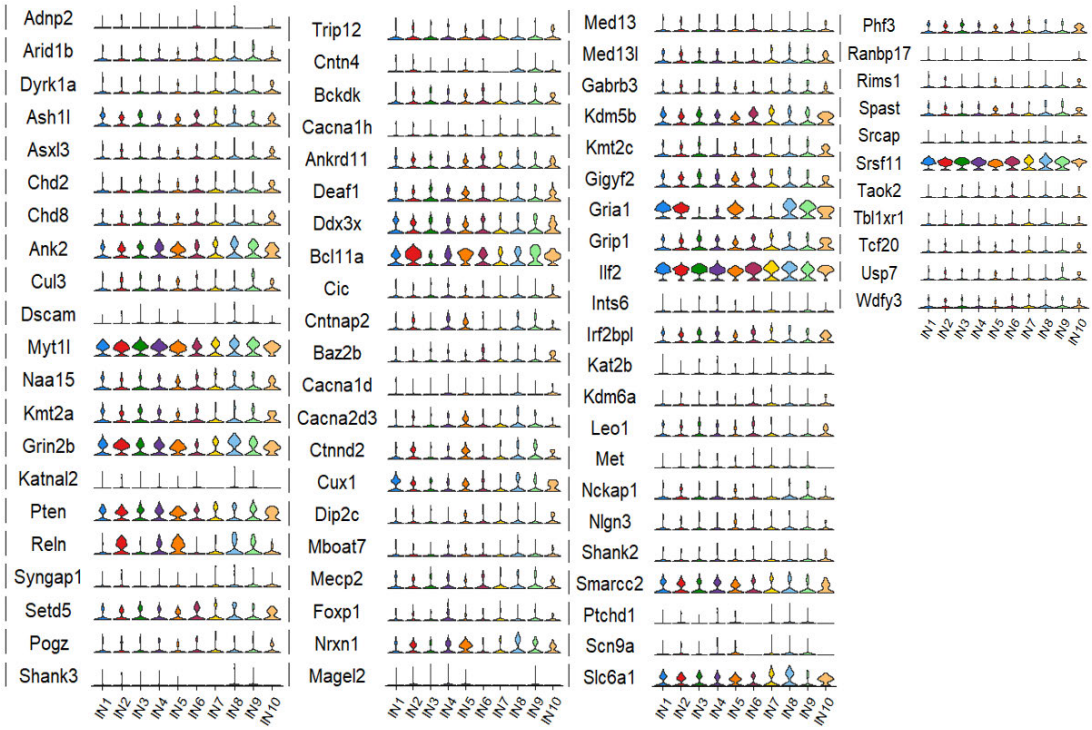


Figure 73: Violin plot illustrating expression pattern of monogenic ASD risk genes across ten interneuron cell clusters in mouse cortex.

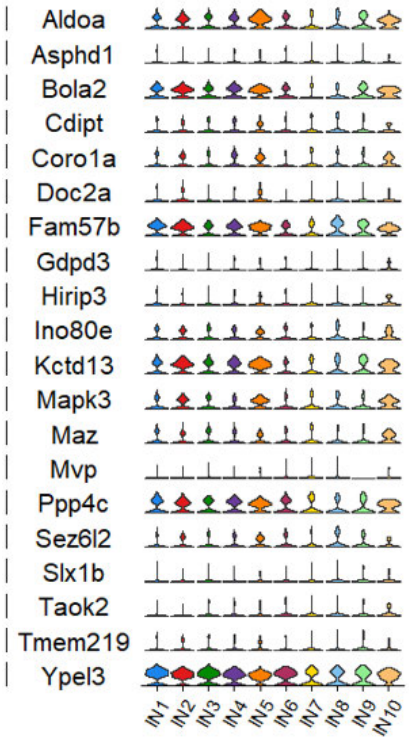


Figure 74: Violin plot illustrating expression pattern of ASD risk genes on *16p11.2* locus across ten interneruon cell clusters in mouse cortex.

5.5 Conclusion

We applied scRNA-seq analysis to identify the diversity of progenitor and interneruons in the embryonic GE regions and reveal the expression pattern of ASD risk genes. There was no significantly differentially expressed ASD risk genes identified across the clusters in E12.5 mouse progenitors. Three monogenic genes (*Med13l, Dyrk1a* and *Bcl11a*) and two genes on *16p11.2* locus (*Aldoa* and *Ypel3*) were significantly differentially expressed across the clusters in E14.5 mouse progenitors. But from the heatmap and violin plots of these genes, we noticed that the expression levels of *Med13l, Dyrk1a* and *Bcl11a* genes were high in several clusters even they were statistically enriched in P5 or P6. So it was hard to judge the possible roles of these genes. The expression levels of *Aldoa* gene were much higher in VZ progenitors (P1), so we hyposised that Aldoa protein may be play a role on cell proliferation. *Ypel3* gene was highly expressed in SVZ progenitors (P11), and previouse study indicated that Ypel3 protein involved in proliferation and apoptosis in myeloid precursor cells.

The unsupervised clustering of mouse interneruons in GE regions shown that the emerging signature of the two original regions of cortical interneurons across several clusters. We could identified the transcriptomic difference between MGE-derived cells (IN13, IN4, IN3, IN1, IN2 and IN9) and CGE-derived cells (IN10, IN5 and IN12). But there were some clusters (IN8, IN7, IN11 and IN6) that express neither MGE or CGE marker genes. There were thirteen monogenic ASD risk genes and one gene on *16p11.2* locus significantly differential expressed across the thirteen clusters. Some of these differential expressed ASD risk genes enriched in MGE-derived cell clusters (IN13, IN3 and IN1), and some of these differential expressed ASD risk genes enriched in CGE-derived cell clusters (IN5 and IN12). This result also indicated that the enrichment of ASD risk genes was depend on the characteristic of interneurons, but not any specific inteneuron lineages.

The cortical interneruons are more maturing than the interneurons in GE region. We characterized the diversity of cortical interneurons generated from E18.5 mouse cortex, and generated ten transcriptionally distinct cell clusters. We identified two of the mouse interneruon clusters (M_IN8 and M_IN9) were remarkably similar to two human interneruon clusters (IN5 and IN8). Many ASD risk genes, for example, eight monogenic genes (*Scn9a, Ctnnd2, Cntnap2, Gria1, Reln, Bcl11a, Cacna2d3* and *Nrxn1*) and three genes on *16p11.2* locus (*Kctd13, Aldoa* and *Doc2a*) were highly expressed in M_IN8 and M_IN9. As we described in Chapter 3, *SCN9A* and *NRXN1* genes were high expressed in human IN5 and IN8, and *GRIA1* genes were enriched in the whole human interneruons. This left five monogenic genes (*Ctnnd2, Cntnap2, Reln, Bcl11a* and *Cacna2d3*) that identified in mouse interneeuron clusters were not differential expressed among human interneruon clusters.

We also noticed that *Reln, Gria1, Cacna2d3,* and *Nrxn1* genes were significantly enriched in Sst+ interneruons. *Cntnap2* and *Cux1* were significantly enriched in Pvalb+ interneruons. And *Bcl11a* gene were significantly enriched in both Sst+ and Pvalb+ interneruons.

To the end, we used a publicly available scRNA-seq dataset of 766 interneurons from the adult mouse visual cortex (Tasic *et al.*, 2018) and identified highly variable genes shared between the adult and embryonic datasets. Firstly, we employed the resulting dataset to identify the features that best represent each of the interneuron cell types found in the adult mouse cortex. We identified eleven interneuron cell types in the embryonic GE regions, and there was not significantly differential expressed ASD risk genes across these cell types. We further combined the eleven interneuron cell types into four cardinal interneuron classes (Sst, Pval, Vip and Ndnf), and confirmed no significant enrichment of ASD risk genes was found in the well-known cardinal interneuron classes.

Chapter 6: General discussion

6.1 Concluding remarks

ASD is highly heritable but genetically heterogeneous. It has been reported that ASD risk genes form co-expression networks or gene sets that are expressed at relatively higher levels in specific cell types, such as interneurons and pyramidal neurons (Skene and Grant, 2016b; Skene *et al.*, 2018; Wang *et al.*, 2018). It also has been suggested that the differences in the expression of ASD risk genes may underlie some critical differences in the organization of inhibitory circuits in humans (Hashemi *et al.*, 2016; Zerbi *et al.*, 2018). But our understanding of the gene expression pattern of ASD risk genes among different cell types during human early cortical development is still very limited.

The recent advent of scRNA-seq has provided biologists with a powerful new tool to gain insight into the developing brain by simultaneously analysing the transcriptomes of thousands of individual cells harvested from developing brain tissue. In the thesis, we take advantage of an scRNA-seq dataset acquired from developing human fetal cortex at a range of gestational stages to investigate which classes are likely to be vulnerable to autism causing mutations by examining the expression of genes associated with ASD with monogenic (the 86 high confidence and strong candidate genes) and polygenetic (the 29 16p11.2 genes) in developing human cortex. We found that 24 ASD risk genes, including 17 monogenic genes and 7 genes on 16p11.2 locus, were significantly differentially expressed across the cardinal cell classes (Figure 12B). We also investigated the expression pattern of ASD risk genes in subclasses of cells comprising the cell cardinal classes and found a pattern of significantly differentially expressed among cell clusters in each cardinal cell class (Figure 17D for NPCs, Figure 21D for ExNs and Figure 24C for INs) with strikingly enriched expression in specific subclasses of interneuron we called IN8.

Based on the investigation of differentially expressed ASD risk genes across cardinal cell classes and cell types there are three major findings in this thesis.

We found, in agreement with other studies, that NPCs, pyramidal neurons and interneurons during human development are potentially vulnerable for ASD since ASD risk genes are highly expressed among them (Skene and Grant, 2016b; Skene *et al.*, 2018; Wang *et al.*, 2018; Griesi-Oliveira *et al.*, 2020). Interneurons were regarded as a disproportionately vulnerable cell class for ASD. This is because most of differentially expressed ASD risk genes were enriched in interneurons (Fig 12B). We noticed that the specific function of some individual ASD risk genes are reported previously. For example, *KIF22* gene, which enriched in NPCs, was reported as a control gene of cell cycle in human cancer cell proliferation (Yu *et al.*, 2014). Other ASD risk genes are categorized as gene sets in previous reports. From this we propose the hypothesis that some ASD risk mutations may affect cell proliferation

The most important novel finding in this thesis is that, within interneurons, the majority of ASD risk genes are highly expressed within a small cluster of developing interneurons, notably IN8. The ASD risk genes that significantly highly expressed in IN8 interneurons can be separated into two large categories, “synapse” and “ion channel”. For example, *CNTN4*, *SHANK2, TRIO* and *GRIA1* genes are related with the development of synapse (Esselmann *et al.*, 2017; Andreae and Burrone, 2018; Heise *et al.*, 2018; Kim *et al.*, 2018; Schidlitzki *et al.*, 2020). *SCN9A*, *CACNA1L* and *ANK2* genes are related with “ion channel”, especially sodium and calcium channel (Perez-Reyes, 2003; Drenth and Waxman, 2007; Meisler, O’Brien and Sharkey, 2010; Kline *et al.*, 2014). From this we propose the hypothesis that many ASD causing mutations primarily affect the electrophysiological function of IN8 interneurons mediated by ion channels and neurotransmitters and therefore change the functionality of neural circuitry so as to predispose to autism. We used a list of well-known marker genes to test if any IN cluster fit into any known interneuron cell type. However, the expression pattern of these marker genes is not clear enough to identify what cell types that these cell cluster were

(Figure 27). Further work is needed to investigate the function of IN8 and how they contribute to brain function.

An important finding of this thesis is that developing mouse brain contains a molecular correlate of IN8. The expression pattern of highly expressed genes is very similar between mouse IN2/5 and human IN5/8 (Figure 71C and D). Also, the expression pattern of part of ASD risk genes are similar between the human interneuron cluster IN8 and the corresponded mouse interneuron cluster IN2/5 (Figure 72A). Both in human and mouse datasets, the expression pattern would validate the same conclusion that the interneuron cell clusters, which enriched ASD risk genes, do not correspond to any well-known interneuron cell type or lineage, but may represent a cell state during their development. The similarity between human and mouse interneuron clusters indicate that we can use rodent models to investigate the function of IN8 during interneuron development.

6.2 Future Work

The advent of single-cell sequencing has enabled the unbiased analysis of molecular profiles of individual cells, highlighting a remarkable level of heterogeneity in cellular populations that may contribute to the phenotypic heterogeneity in disease (Polioudakis *et al.*, 2018; Skene *et al.*, 2018). In this thesis, we have used a set of published scRNA-seq datasets to map ASD risk genes to specific cell types in human and mouse fetal cortex, implicating dysregulation of specific cell types, as the mechanistic underpinnings of the ASD. However, more works need to be done to test the three hypotheses above.

To test the hypothesis that the mutation of genes whose expression is enriched in IN8 affects the function on interneurons and how the IN8 properties altered in human/mouse mutants, we can perform experiments on rodents and human

iPSC ASD models harbouring gain or loss of function mutations recapitulating autism patient genotypes (Muotri, 2016; Vitrac and Cloëz-Tayarani, 2018; Grunwald *et al.*, 2019; Gordon and Geschwind, 2020). Based on our findings we would first specifically investigate altered electrophysiological properties of IN8 interneurons as predicted by our hypothesis. The mouse mutants with similar genotype to human patients can be used to explore the genetic function of heterozygous mutation of the ASD monogenic risk genes. For example, the CRISPR/Cas9 methods can be used to introduce mutations and control genetic inheritance in rodent model or iPSC cells (Bassett, 2017; Powell *et al.*, 2017). The electrophysiological experiment can be performed by multi-electrode array analysis of iPSC-derived neurons or mouse brain slices (Kazdoba *et al.*, 2016; Deshpande *et al.*, 2017).

For the second hypothesis, we aim to extend our understanding of what is IN8. We plan to link transcriptional profiles of ASD risk genes to diversity of cell types during the sequential specification of human cortical interneuron development. The human datasets we used in this thesis can only cover the developmental stages from GW08 to GW26 (Figure 47A), the stage at which IN8 started to appear. So, the cells collected from the developmental stages later than GW26 will be important to characterise subsequent development of IN8 to gain insight into how they might be affected by Autism causing mutations. We also plan to perform the scRNA-seq analyses on snap-frozen brain samples from adult patients with ASD, and compare with brain samples from neurologically normal age-and sex-matched controls. By performing unbiased clustering of cell types based on single-cell transcriptional profiles and comparing ASD risk gene expression in each cell type between autism and control groups, we expect to find out the similar cell cluster as the human IN8 in the control human data. We will also combine the control and autism datasets together to analysis the correlation between the cell clusters in control and autism, and we will compare the expression pattern of ASD risk genes among the cell clusters, as well as between control and autism conditions.

The exact nature of IN8 in rodents requires further investigation, not only to determine how good a model rodent is for studying this cell type but also to understand the evolution of these cells in humans. This can be tacked by employing scRNA-seq data sets from both mouse and human spanning more developmental stages and by using *in-situ* hybridizations for the conserved marker genes between human IN8 and mouse IN2/5 on tissue sections to understand more about their cell biology in tractable rodent models. Recently, we noticed that there were some studies claimed that the cell type between human and chimpanzee are very similar (Marchetto *et al.*, 2019; Khrameeva *et al.*, 2020). So multiple species maybe worth to be included in further analysis to test the hypothesis about if the properties of IN8-like interneurons altered in the different animal models.

To conclude, we will extend out our analysis of the novel cell-state IN8, investigate its role in the development of autism and study the roles of ASD risk genes likely to be important for these cells from the single-cell sequencing analyses in this thesis. The scRNA-seq datasets that cover the cells from early fetal stages to adult could give us a dynamical system view of the human cortical development. And based on these datasets, we can answer more questions about the IN8, such as where will these cells end up. The comparison between control and autism human brains could tell us what happens to IN8 cells as they differentiate in the autism state. Moreover, the correlated cell clusters of human IN8, which we found in mouse dataset as IN2/5, give us confidence to use rodent models to investigate how the ASD risk gens maybe trigger the autism. The wet lab experiments, such as electrophysiological experiments, can be used not only in embryonic human brain slices, but also on the rodent mutant models, to explore the biological function of the enriched ASD risk genes in the human IN8 cell cluster.

References:

Abriel, H. and Kass, R. S. (2005) ‘Regulation of the voltage-gated cardiac sodium channel Nav1.5 by interacting proteins’, *Trends in Cardiovascular Medicine*, pp. 35–40. doi: 10.1016/j.tcm.2005.01.001.

Van den Ameele, J. *et al.* (2014) ‘Thinking out of the dish: What to learn about cortical development using pluripotent stem cells’, *Trends in Neurosciences*. Elsevier Ltd, pp. 334–342. doi: 10.1016/j.tins.2014.03.005.

Anderson, S. A. (2001) *Subcortical origins of cortical interneurons*.

Andreae, L. C. and Burrone, J. (2018) ‘The role of spontaneous neurotransmission in synapse and circuit development’, *Journal of Neuroscience Research*. John Wiley and Sons Inc., 96(3), pp. 354–359. doi: 10.1002/jnr.24154.

Ascoli, G. A. *et al.* (2008) ‘Petilla terminology: Nomenclature of features of GABAergic interneurons of the cerebral cortex’, *Nature Reviews Neuroscience*, pp. 557–568. doi: 10.1038/nrn2402.

Ba, W. *et al.* (2016) ‘TRIO loss of function is associated with mild intellectual disability and affects dendritic branching and synapse function’, *Human Molecular Genetics*, 25(5), pp. 892–902. doi: 10.1093/hmg/ddv618.

Banerjee-Basu, S. and Packer, A. (2010) ‘SFARI Gene: An evolving database for the autism research community’, *DMM*

Disease Models and Mechanisms, pp. 133–135. doi: 10.1242/dmm.005439.

Bartolini, G., Ciceri, G. and Marín, O. (2013) 'Integration of GABAergic Interneurons into Cortical Cell Assemblies: Lessons from Embryos and Adults', *Neuron*, pp. 849–864. doi: 10.1016/j.neuron.2013.08.014.

Bassett, A. R. (2017) 'Editing the genome of hiPSC with CRISPR/Cas9: disease models', *Mammalian Genome*. Springer New York LLC, pp. 348–364. doi: 10.1007/s00335-017-9684-9.

Bergles, D. E. and Richardson, W. D. (2016) 'Oligodendrocyte development and plasticity', *Cold Spring Harbor Perspectives in Biology*. Cold Spring Harbor Laboratory Press, 8(2). doi: 10.1101/cshperspect.a020453.

Blaker-Lee, A. *et al.* (2012) 'Zebrafish homologs of genes within 16p11.2, a genomic region associated with brain disorders, are active during brain development, and include two deletion dosage sensor genes', *DMM Disease Models and Mechanisms*, 5(6), pp. 834–851. doi: 10.1242/dmm.009944.

Braccioli, L. *et al.* (2017) 'FOXP1 Promotes Embryonic Neural Stem Cell Differentiation by Repressing Jagged1 Expression', *Stem Cell Reports*. Cell Press, 9(5), pp. 1530–1545. doi: 10.1016/j.stemcr.2017.10.012.

Butler, A. *et al.* (2018) 'Integrating single-cell transcriptomic data across different conditions, technologies, and species', *Nature Biotechnology*. Nature Publishing Group, 36(5), pp. 411–420. doi:

10.1038/nbt.4096.

Butt, S. J. B. *et al.* (2005) ‘The temporal and spatial origins of cortical interneurons predict their physiological subtype’, *Neuron*, 48(4), pp. 591–604. doi: 10.1016/j.neuron.2005.09.034.

Cai, Y. *et al.* (2013) ‘Nuclear receptor COUP-TFII-expressing neocortical interneurons are derived from the medial and lateral/caudal ganglionic eminence and define specific subsets of mature interneurons’, *Journal of Comparative Neurology*, 521(2), pp. 479–497. doi: 10.1002/cne.23186.

Camp, J. G. *et al.* (2015) ‘Human cerebral organoids recapitulate gene expression programs of fetal neocortex development’, *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 112(51), pp. 15672–15677. doi: 10.1073/pnas.1520760112.

Chen, J. G. *et al.* (2005) ‘Zfp312 is required for subcortical axonal projections and dendritic morphology of deep-layer pyramidal neurons of the cerebral cortex’, *Proceedings of the National Academy of Sciences of the United States of America*, 102(49), pp. 17792–17797. doi: 10.1073/pnas.0509032102.

Chen, X. *et al.* (2008) ‘Integration of External Signaling Pathways with the Core Transcriptional Network in Embryonic Stem Cells’, *Cell*, 133(6), pp. 1106–1117. doi: 10.1016/j.cell.2008.04.043.

Chung, K. M. *et al.* (2012) ‘Cross cultural differences in challenging behaviors of children with autism spectrum disorders: An international examination between Israel, South Korea, the United

Kingdom, and the United States of America', *Research in Autism Spectrum Disorders*, 6(2), pp. 881–889. doi: 10.1016/j.rasd.2011.03.016.

Cobos, I. *et al.* (2005) 'Mice lacking Dlx1 show subtype-specific loss of interneurons, reduced inhibition and epilepsy', *Nature Neuroscience*, 8(8), pp. 1059–1068. doi: 10.1038/nn1499.

Colasante, G. *et al.* (2008) 'Arx is a direct target of Dlx2 and thereby contributes to the tangential migration of GABAergic interneurons', *Journal of Neuroscience*, 28(42), pp. 10674–10686. doi: 10.1523/JNEUROSCI.1283-08.2008.

Constantino, J. N. and Charman, T. (2016) 'Diagnosis of autism spectrum disorder: reconciling the syndrome, its diverse origins, and variation in expression', *The Lancet Neurology*. Lancet Publishing Group, pp. 279–291. doi: 10.1016/S1474-4422(15)00151-9.

Costa, M. R. and Müller, U. (2015) 'Specification of excitatory neurons in the developing cerebral cortex: Progenitor diversity and environmental influences', *Frontiers in Cellular Neuroscience*. Frontiers Media S.A., pp. 1–9. doi: 10.3389/fncel.2014.00449.

Crow, M. *et al.* (2018) 'Characterizing the replicability of cell types defined by single cell RNA-sequencing data using MetaNeighbor', *Nature Communications*. Nature Publishing Group, 9(1). doi: 10.1038/s41467-018-03282-0.

Darmanis, S. *et al.* (2015) 'A survey of human brain transcriptome diversity at the single cell level', *Proceedings of the National*

Academy of Sciences of the United States of America. National Academy of Sciences, 112(23), pp. 7285–7290. doi: 10.1073/pnas.1507125112.

Dehay, C., Kennedy, H. and Kosik, K. S. (2015) 'The Outer Subventricular Zone and Primate-Specific Cortical Complexification', *Neuron*. Cell Press, pp. 683–694. doi: 10.1016/j.neuron.2014.12.060.

Deshpande, A. *et al.* (2017) 'Cellular Phenotypes in Human iPSC-Derived Neurons from a Genetic Model of Autism Spectrum Disorder', *Cell Reports*. Elsevier B.V., 21(10), pp. 2678–2687. doi: 10.1016/j.celrep.2017.11.037.

Devor, A. *et al.* (2017) 'Genetic evidence for role of integration of fast and slow neurotransmission in schizophrenia HHS Public Access', *Mol Psychiatry*, 22(6), pp. 792–801. doi: 10.1038/mp.2017.33.

Ding, L. *et al.* (2015) 'Systems Analyses Reveal Shared and Diverse Attributes of Oct4 Regulation in Pluripotent Cells', *Cell Systems*. Cell Press, 1(2), pp. 141–151. doi: 10.1016/j.cels.2015.08.002.

Drenth, J. P. H. and Waxman, S. G. (2007) 'Mutations in sodium-channel gene SCN9A cause a spectrum of human genetic pain disorders', *Journal of Clinical Investigation*. American Society for Clinical Investigation, pp. 3603–3609. doi: 10.1172/JCI33297.

Du, T. *et al.* (2008) 'NKX2.1 specifies cortical interneuron fate by activating Lhx6', *Development*, 135(8), pp. 1559–1567. doi:

10.1242/dev.015123.

Elsen, G. E. *et al.* (2018) ‘The epigenetic factor landscape of developing neocortex is regulated by transcription factors Pax6→ Tbr2→ Tbr1’, *Frontiers in Neuroscience*. Frontiers Media S.A., 12(AUG). doi: 10.3389/fnins.2018.00571.

Esselmann, S. G. A. *et al.* (2017) ‘Synaptic homeostasis requires the membrane-proximal carboxy tail of GluA2’, *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 114(50), pp. 13266–13271. doi: 10.1073/pnas.1716022114.

Fan, X. *et al.* (2018) ‘Spatial transcriptomic survey of human embryonic cerebral cortex by single-cell RNA-seq analysis’, *Cell Research*. Nature Publishing Group, 28(7), pp. 730–745. doi: 10.1038/s41422-018-0053-3.

Fertuzinhos, S. *et al.* (2014) ‘Laminar and temporal expression dynamics of coding and noncoding RNAs in the mouse neocortex’, *Cell Reports*. Elsevier, 6(5), pp. 938–950. doi: 10.1016/j.celrep.2014.01.036.

Gao, P. *et al.* (2013) ‘Lineage-dependent circuit assembly in the neocortex’, *Development (Cambridge)*, pp. 2645–2655. doi: 10.1242/dev.087668.

Gao, Z. *et al.* (2011) ‘The master negative regulator REST/NRSF controls adult neurogenesis by restraining the neurogenic program in quiescent stem cells’, *Journal of Neuroscience*, 31(26), pp. 9772–9786. doi: 10.1523/JNEUROSCI.1604-11.2011.

Gelman, D. *et al.* (2011) ‘A wide diversity of cortical GABAergic interneurons derives from the embryonic preoptic area’, *Journal of Neuroscience*, 31(46), pp. 16570–16580. doi: 10.1523/JNEUROSCI.4068-11.2011.

Gelman, D. M. and Marín, O. (2010) ‘Generation of interneuron diversity in the mouse cerebral cortex’, *European Journal of Neuroscience*, pp. 2136–2141. doi: 10.1111/j.1460-9568.2010.07267.x.

Gordon, A. and Geschwind, D. H. (2020) ‘Human in vitro models for understanding mechanisms of autism spectrum disorder’, *Molecular Autism*. BioMed Central Ltd., pp. 1–18. doi: 10.1186/s13229-020-00332-7.

Greig, L. C. *et al.* (2013) ‘Molecular logic of neocortical projection neuron specification, development and diversity’, *Nature Reviews Neuroscience*, pp. 755–769. doi: 10.1038/nrn3586.

Griesi-Oliveira, K. *et al.* (2020) ‘Transcriptome of iPSC-derived neuronal cells reveals a module of co-expressed genes consistently associated with autism spectrum disorder’, *Molecular Psychiatry*. Springer Nature, pp. 1–17. doi: 10.1038/s41380-020-0669-9.

Grün, D. *et al.* (2015) ‘Single-cell messenger RNA sequencing reveals rare intestinal cell types’, *Nature*. Nature Publishing Group, 525(7568), pp. 251–255. doi: 10.1038/nature14966.

Grunwald, L. M. *et al.* (2019) ‘Comparative characterization of human induced pluripotent stem cells (hiPSC) derived from patients

with schizophrenia and autism', *Translational Psychiatry*. Nature Publishing Group, 9(1), pp. 1–11. doi: 10.1038/s41398-019-0517-3.

Hansen, D. V. *et al.* (2010) 'Neurogenic radial glia in the outer subventricular zone of human neocortex', *Nature*, 464(7288), pp. 554–561. doi: 10.1038/nature08845.

Hashemi, E. *et al.* (2016) 'The Number of Parvalbumin-Expressing Interneurons Is Decreased in the Prefrontal Cortex in Autism'. doi: 10.1093/cercor/bhw021.

Heise, C. *et al.* (2018) 'Heterogeneity of Cell Surface Glutamate and GABA Receptor Expression in Shank and CNTN4 Autism Mouse Models', *Frontiers in Molecular Neuroscience*. Frontiers Media S.A., 11, p. 212. doi: 10.3389/fnmol.2018.00212.

Hendry, S. H. *et al.* (1987) 'Numbers and proportions of GABA-immunoreactive neurons in different areas of monkey cerebral cortex.', *Journal of Neuroscience*, 7(5), pp. 1503–1519. doi: 10.1523/jneurosci.07-05-01503.1987.

Hodge, R. D. *et al.* (2018) 'Conserved cell types with divergent features between human and mouse cortex', *bioRxiv*, p. 384826. doi: 10.1101/384826.

Huber, W. *et al.* (2015) 'Orchestrating high-throughput genomic analysis with Bioconductor', *Nature Methods*. Nature Publishing Group, 12(2), pp. 115–121. doi: 10.1038/nmeth.3252.

Inan, M., Welagen, J. and Anderson, S. A. (2012) 'Spatial and

temporal bias in the mitotic origins of somatostatin- and parvalbumin-expressing interneuron subgroups and the chandelier subtype in the medial ganglionic eminence', *Cerebral Cortex*, 22(4), pp. 820–827. doi: 10.1093/cercor/bhr148.

Jakovcevski, M. *et al.* (2015) 'Neuronal Kmt2a/Mll1 histone methyltransferase is essential for prefrontal synaptic plasticity and working memory', *Journal of Neuroscience*. Society for Neuroscience, 35(13), pp. 5097–5108. doi: 10.1523/JNEUROSCI.3004-14.2015.

Kageyama, J. *et al.* (2018) 'ShinyCortex: Exploring Single-Cell Transcriptome Data From the Developing Human Cortex', *Frontiers in Neuroscience*, 12. doi: 10.3389/fnins.2018.00315.

Kanatani, S. *et al.* (2008) 'COUP-TFII is preferentially expressed in the caudal ganglionic eminence and is involved in the caudal migratory stream', *Journal of Neuroscience*, 28(50), pp. 13582–13591. doi: 10.1523/JNEUROSCI.2132-08.2008.

Kang, H. J. *et al.* (2011) 'Spatio-temporal transcriptome of the human brain', *Nature*, 478(7370), pp. 483–489. doi: 10.1038/nature10523.

Kazdoba, T. M. *et al.* (2016) 'Translational mouse models of autism: Advancing toward pharmacological therapeutics', in *Current Topics in Behavioral Neurosciences*. Springer Verlag, pp. 1–52. doi: 10.1007/7854_2015_5003.

Kelsom, C. and Lu, W. (2013) 'Development and specification of GABAergic cortical interneurons', *Cell and Bioscience*. doi:

10.1186/2045-3701-3-19.

Kepecs, A. and Fishell, G. (2014) ‘Interneuron cell types are fit to function’, *Nature*, pp. 318–326. doi: 10.1038/nature12983.

Khrameeva, E. *et al.* (2020) ‘Single-cell-resolution transcriptome map of human, chimpanzee, bonobo, and macaque brains’, *Genome Research*. Cold Spring Harbor Laboratory Press, 30(5), pp. 776–789. doi: 10.1101/gr.256958.119.

Kim, R. *et al.* (2018) ‘Cell-type-specific shank2 deletion in mice leads to differential synaptic and behavioral phenotypes’, *Journal of Neuroscience*. Society for Neuroscience, 38(17), pp. 4076–4092. doi: 10.1523/JNEUROSCI.2684-17.2018.

Klassen, T. *et al.* (2011) ‘Exome sequencing of ion channel genes reveals complex profiles confounding personal risk assessment in epilepsy’, *Cell*. NIH Public Access, 145(7), pp. 1036–1048. doi: 10.1016/j.cell.2011.05.025.

Kline, C. F. *et al.* (2014) ‘Ankyrin-B regulates Cav2.1 and Cav2.2 channel expression and targeting’, *Journal of Biological Chemistry*. American Society for Biochemistry and Molecular Biology Inc., 289(8), pp. 5285–5295. doi: 10.1074/jbc.M113.523639.

Kriegstein, A., Noctor, S. and Martínez-Cerdeño, V. (2006) ‘Patterns of neural stem and progenitor cell division may underlie evolutionary cortical expansion’, *Nature Reviews Neuroscience*, pp. 883–890. doi: 10.1038/nrn2008.

Kwan, K. Y., Šestan, N. and Anton, E. S. (2012) ‘Transcriptional co-

regulation of neuronal migration and laminar identity in the neocortex', *Development*, pp. 1535–1546. doi: 10.1242/dev.069963.

De La Torre-Ubieta, L. *et al.* (2016) 'Advancing the understanding of autism disease mechanisms through genetics', *Nature Medicine*. Nature Publishing Group, pp. 345–361. doi: 10.1038/nm.4071.

Lake, B. B. *et al.* (2016) 'Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain', *Science*. American Association for the Advancement of Science, 352(6293), pp. 1586–1590. doi: 10.1126/science.aaf1204.

Lewis, D. A. *et al.* (2012) 'Cortical parvalbumin interneurons and cognitive dysfunction in schizophrenia', *Trends in Neurosciences*, pp. 57–67. doi: 10.1016/j.tins.2011.10.004.

Lewitus, E., Kelava, I. and Huttner, W. B. (2013) 'Conical expansion of the outer subventricular zone and the role of neocortical folding in evolution and development', *Frontiers in Human Neuroscience*. Frontiers Media S. A., (JUL). doi: 10.3389/fnhum.2013.00424.

Lin, G. N. *et al.* (2015) 'Spatiotemporal 16p11.2 Protein Network Implicates Cortical Late Mid-Fetal Brain Development and KCTD13-Cul3-RhoA Pathway in Psychiatric Diseases', *Neuron*. Cell Press, 85(4), pp. 742–754. doi: 10.1016/j.neuron.2015.01.010.

Lin, L. C. and Sibille, E. (2013) 'Reduced brain somatostatin in mood disorders: A common pathophysiological substrate and drug target?', *Frontiers in Pharmacology*. doi: 10.3389/fphar.2013.00110.

Lodato, S. *et al.* (2011) ‘Excitatory Projection Neuron Subtypes Control the Distribution of Local Inhibitory Interneurons in the Cerebral Cortex’, *Neuron*, 69(4), pp. 763–779. doi: 10.1016/j.neuron.2011.01.015.

Loo, L. *et al.* (2019) ‘Single-cell transcriptomic analysis of mouse neocortical development’, *Nature Communications*. Nature Publishing Group, 10(1). doi: 10.1038/s41467-018-08079-9.

Van Der Maaten, L. and Weinberger, K. (2012) *STOCHASTIC TRIPLET EMBEDDING*, *IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING*.

Machol, K. *et al.* (2019) ‘Expanding the Spectrum of BAF-Related Disorders: De Novo Variants in SMARCC2 Cause a Syndrome with Intellectual Disability and Developmental Delay’, *American Journal of Human Genetics*. Cell Press, 104(1), pp. 164–178. doi: 10.1016/j.ajhg.2018.11.007.

Marchetto, M. C. *et al.* (2019) ‘Species-specific maturation profiles of human, chimpanzee and bonobo neural cells’, *eLife*. eLife Sciences Publications Ltd, 8. doi: 10.7554/eLife.37527.

Mattison, K. A. *et al.* (2018) ‘*SLC6A1* variants identified in epilepsy patients reduce γ-aminobutyric acid transport’, *Epilepsia*. Blackwell Publishing Inc., 59(9), pp. e135–e141. doi: 10.1111/epi.14531.

Mayer, C. *et al.* (2018) ‘Developmental diversification of cortical inhibitory interneurons’, *Nature*. Nature Publishing Group, 555(7697), pp. 457–462. doi: 10.1038/nature25999.

Meisler, M. H., O'Brien, J. E. and Sharkey, L. M. (2010) 'Sodium channel gene family: Epilepsy mutations, gene interactions and modifier effects', *Journal of Physiology*, pp. 1841–1848. doi: 10.1113/jphysiol.2010.188482.

Melzer, S. *et al.* (2017) 'Distinct Corticostriatal GABAergic Neurons Modulate Striatal Output Neurons and Motor Activity', *CellReports*, 19, pp. 1045–1055. doi: 10.1016/j.celrep.2017.04.024.

Menassa, D. A. and Gomez-Nicola, D. (2018) 'Microglial Dynamics During Human Brain Development', *Frontiers in Immunology*, 9. doi: 10.3389/fimmu.2018.01014.

Mi, D. *et al.* (2018) 'Early emergence of cortical interneuron diversity in the mouse embryo', *Science*. American Association for the Advancement of Science, 360(6384), pp. 81–85. doi: 10.1126/science.aar6821.

Miyoshi, G. *et al.* (2015) 'Prox1 regulates the subtype-specific development of caudal ganglionic eminence-derived GABAergic cortical interneurons', *Journal of Neuroscience*. Society for Neuroscience, 35(37), pp. 12869–12889. doi: 10.1523/JNEUROSCI.1164-15.2015.

Morson, S. *et al.* (2019) 'Expression of genes in the 16p11.2 locus during human fetal cortical neurogenesis'. doi: 10.1101/633461.

Mozhui, K. *et al.* (2011) 'Genetic regulation of Nrnx1 expression: An integrative cross-species analysis of schizophrenia candidate genes', *Translational Psychiatry*. Nature Publishing Group, 1. doi: 10.1038/tp.2011.24.

Mullins, J. L., Chung, S. K. and Rees, M. I. (2010) ‘Fine architecture and mutation mapping of human brain inhibitory system ligand gated ion channels by high-throughput homology modeling’, in *Advances in Protein Chemistry and Structural Biology*. Academic Press Inc., pp. 117–152. doi: 10.1016/B978-0-12-381264-3.00004-7.

Muotri, A. R. (2016) ‘The human model: Changing focus on autism research’, *Biological Psychiatry*. Elsevier USA, pp. 642–649. doi: 10.1016/j.biopsych.2015.03.012.

Nery, S., Fishell, G. and Corbin, J. G. (2002) ‘The caudal ganglionic eminence is a source of distinct cortical and subcortical cell populations’, *Nature Neuroscience*, 5(12), pp. 1279–1287. doi: 10.1038/nn971.

Niquille, M. *et al.* (2018) ‘Neurogliaform cortical interneurons derive from cells in the preoptic area’, *eLife*. eLife Sciences Publications Ltd, 7. doi: 10.7554/eLife.32017.

Nowakowski, T. J. *et al.* (2016) ‘Transformation of the Radial Glia Scaffold Demarcates Two Stages of Human Cerebral Cortex Development’, *Neuron*. Cell Press, 91(6), pp. 1219–1227. doi: 10.1016/j.neuron.2016.09.005.

Nowakowski, T. J. *et al.* (2017) ‘Spatiotemporal gene expression trajectories reveal developmental hierarchies of the human cortex’, *Science*. American Association for the Advancement of Science, 358(6368), pp. 1318–1323. doi: 10.1126/science.aap8809.

Ohtaka-Maruyama, C. and Okado, H. (2015) ‘Molecular pathways

underlying projection neuron production and migration during cerebral cortical development', *Frontiers in Neuroscience*. Frontiers Media S.A. doi: 10.3389/fnins.2015.00447.

Olson, E. C. (2014) 'Analysis of preplate splitting and early cortical development illuminates the biology of neurological disease', *Frontiers in Pediatrics*. Frontiers Media S.A. doi: 10.3389/fped.2014.00121.

Packer, A. (2016) 'Neocortical neurogenesis and the etiology of autism spectrum disorder', *Neuroscience and Biobehavioral Reviews*. Elsevier Ltd, pp. 185–195. doi: 10.1016/j.neubiorev.2016.03.002.

Parikshak, N. N. *et al.* (2013) 'XIntegrative functional genomic analyses implicate specific molecular pathways and circuits in autism', *Cell*. Cell Press, 155(5), p. 1008. doi: 10.1016/j.cell.2013.10.031.

Peñagarikano, O. *et al.* (2011) 'Absence of CNTNAP2 leads to epilepsy, neuronal migration abnormalities, and core autism-related deficits', *Cell*, 147(1), pp. 235–246. doi: 10.1016/j.cell.2011.08.040.

Perez-Reyes, E. (2003) 'Molecular physiology of low-voltage-activated T-type calcium channels', *Physiological Reviews*. American Physiological Society, pp. 117–161. doi: 10.1152/physrev.00018.2002.

Philippe, T. J., Tristan, © and Philippe, J. (no date) *Deaf1 and MeCP2 interact to coordinately regulate 5-HT 1A receptor gene*

expression. Available at: http://beta.biomedcentral.com/about/policies/reprints-and-permissions (Accessed: 5 November 2019).

Polioudakis, D. *et al.* (2018) ‘A single cell transcriptomic analysis of human neocortical development’, *bioRxiv*. Cold Spring Harbor Laboratory, p. 401885. doi: 10.1101/401885.

Pollen, A. A. *et al.* (2015) ‘Molecular Identity of Human Outer Radial Glia during Cortical Development’, *Cell*. Cell Press, 163(1), pp. 55–67. doi: 10.1016/j.cell.2015.09.004.

Powell, S. K. *et al.* (2017) ‘Application of CRISPR/Cas9 to the study of brain development and neuropsychiatric disease’, *Molecular and Cellular Neuroscience*. Academic Press Inc., pp. 157–166. doi: 10.1016/j.mcn.2017.05.007.

Pramod, A. B. *et al.* (2013) ‘SLC6 transporters: Structure, function, regulation, disease association and therapeutics’, *Molecular Aspects of Medicine*, pp. 197–219. doi: 10.1016/j.mam.2012.07.002.

Qiu, X. *et al.* (2017) ‘Reversed graph embedding resolves complex single-cell trajectories’, *Nature Methods*. Nature Publishing Group, 14(10), pp. 979–982. doi: 10.1038/nmeth.4402.

Reemst, K. *et al.* (2016) ‘The indispensable roles of microglia and astrocytes during brain development’, *Frontiers in Human Neuroscience*. Frontiers Media S. A, 10(NOV2016). doi: 10.3389/fnhum.2016.00566.

Rossignol, E. (2011) ‘Genetics and function of neocortical GABAergic interneurons in neurodevelopmental disorders’, *Neural Plasticity*. Hindawi Publishing Corporation. doi: 10.1155/2011/649325.

Rubin, A. N. and Kessaris, N. (2013) ‘PROX1: A Lineage Tracer for Cortical Interneurons Originating in the Lateral/Caudal Ganglionic Eminence and Preoptic Area’, *PLoS ONE*. Edited by D. Henrique, 8(10), p. e77339. doi: 10.1371/journal.pone.0077339.

Rudy, B. *et al.* (2011) ‘Three groups of interneurons account for nearly 100% of neocortical GABAergic neurons’, *Developmental Neurobiology*, 71(1), pp. 45–61. doi: 10.1002/dneu.20853.

Saliba, A. E. *et al.* (2014) ‘Single-cell RNA-seq: Advances and future challenges’, *Nucleic Acids Research*. Oxford University Press, pp. 8845–8860. doi: 10.1093/nar/gku555.

Satija, R. *et al.* (2015) ‘Spatial reconstruction of single-cell gene expression data’, *Nature Biotechnology*. Nature Publishing Group, 33(5), pp. 495–502. doi: 10.1038/nbt.3192.

Schidlitzki, A. *et al.* (2020) ‘Proof-of-concept that network pharmacology is effective to modify development of acquired temporal lobe epilepsy’, *Neurobiology of Disease*. Academic Press Inc., 134. doi: 10.1016/j.nbd.2019.104664.

Schoch, S. *et al.* (2002) ‘RIM1α forms a protein scaffold for regulating neurotransmitter release at the active zone’, *Nature*, 415(6869), pp. 321–326. doi: 10.1038/415321a.

Shi, X. *et al.* (2009) ‘Missense mutation of the sodium channel gene SCN2A causes Dravet syndrome’, *Brain and Development*. Brain Dev, 31(10), pp. 758–762. doi: 10.1016/j.braindev.2009.08.009.

Silbereis, J. C. *et al.* (2016) ‘The Cellular and Molecular Landscapes of the Developing Human Central Nervous System’, *Neuron*. Cell Press, p. 248. doi: 10.1016/j.neuron.2015.12.008.

Simunovic, F. *et al.* (no date) ‘Gene expression profiling of substantia nigra dopamine neurons: further insights into Parkinson’s disease pathology’, *A JOURNAL OF NEUROLOGY*. doi: 10.1093/brain/awn323.

Skene, N. G. *et al.* (2018) ‘Genetic identification of brain cell types underlying schizophrenia’, *Nature Genetics*. Nature Publishing Group, 50(6), pp. 825–833. doi: 10.1038/s41588-018-0129-5.

Skene, N. G. and Grant, S. G. N. (2016a) ‘Identification of vulnerable cell types in major brain disorders using single cell transcriptomes and expression weighted cell type enrichment’, *Frontiers in Neuroscience*. Frontiers Media S.A., 10(JAN). doi: 10.3389/fnins.2016.00016.

Skene, N. G. and Grant, S. G. N. (2016b) ‘Identification of vulnerable cell types in major brain disorders using single cell transcriptomes and expression weighted cell type enrichment’, *Frontiers in Neuroscience*. Frontiers Media S.A., 10(JAN), p. 16. doi: 10.3389/fnins.2016.00016.

Smart, I. H. M. (2002) ‘Unique Morphological Features of the Proliferative Zones and Postmitotic Compartments of the Neural

Epithelium Giving Rise to Striate and Extrastriate Cortex in the Monkey', *Cerebral Cortex*. Oxford University Press (OUP), 12(1), pp. 37–53. doi: 10.1093/cercor/12.1.37.

Stanco, A. *et al.* (2014) 'NPAS1 Represses the Generation of Specific Subtypes of Cortical Interneurons', *Neuron*. Cell Press, 84(5), pp. 940–953. doi: 10.1016/j.neuron.2014.10.040.

Sussel, L. (1999) *Forebrain respecification in Nkx2.1 mutants*.

Suuberg, A. (2018) 'Phenotypic and Evolutionary Consequences of Deletion, Duplication, and Triplication at 16p11.2', *SSRN Electronic Journal*. Elsevier BV. doi: 10.2139/ssrn.3185741.

Suzuki, I. K. and Vanderhaeghen, P. (2015) 'Is this a brain which i see before me? Modeling human neural development with pluripotent stem cells', *Development (Cambridge)*. Company of Biologists Ltd, pp. 3138–3150. doi: 10.1242/dev.120568.

Tasic, B. *et al.* (2018) 'Shared and distinct transcriptomic cell types across neocortical areas', *Nature*. Nature Publishing Group, 563(7729), pp. 72–78. doi: 10.1038/s41586-018-0654-5.

Thomsen, C. *et al.* (no date) 'Fixed single-cell transcriptomic characterization of human radial glial diversity'. doi: 10.1038/nmeth.3629.

Vitrac, A. and Cloëz-Tayarani, I. (2018) 'Induced pluripotent stem cells as a tool to study brain circuits in autism-related disorders', *Stem Cell Research and Therapy*. BioMed Central Ltd., p. 226. doi: 10.1186/s13287-018-0966-2.

Volk, D. W. *et al.* (2015) ‘Chemokine receptors and cortical interneuron dysfunction in schizophrenia’, *Schizophrenia Research*. Elsevier, 167(1–3), pp. 12–17. doi: 10.1016/j.schres.2014.10.031.

Wang, P. *et al.* (2018) ‘Enriched expression of genes associated with autism spectrum disorders in human inhibitory neurons’, *Translational Psychiatry*. Nature Publishing Group, 8(1), pp. 1–10. doi: 10.1038/s41398-017-0058-6.

Watson, J. F., Ho, H. and Greger, I. H. (2017) ‘Synaptic transmission and plasticity require AMPA receptor anchoring via its N-terminal domain’, *eLife*. eLife Sciences Publications Ltd, 6. doi: 10.7554/eLife.23024.

Wichterle, H. *et al.* (2001) ‘In utero fate mapping reveals distinct migratory pathways and fates of neurons born in the mammalian basal forebrain’, *Development*, 128(19), pp. 3759–3771.

Wiegreffe, C. *et al.* (2015) ‘Bcl11a (Ctip1) Controls Migration of Cortical Projection Neurons through Regulation of Sema3c’, *Neuron*. Cell Press, 87(2), pp. 311–325. doi: 10.1016/j.neuron.2015.06.023.

Willsey, A. J. *et al.* (2013) ‘XCoexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism’, *Cell*. Cell Press, 155(5), p. 997. doi: 10.1016/j.cell.2013.10.020.

Xu, Q. *et al.* (2004) ‘Journal of Neuroscience’, *J. Neurosci.* Society for Neuroscience, 21(12), pp. 4356–4365. doi: 20026564.

Xu, Q., Tam, M. and Anderson, S. A. (2008) ‘Fate mapping Nkx2.1-lineage cells in the mouse telencephalon’, *Journal of Comparative Neurology*, 506(1), pp. 16–29. doi: 10.1002/cne.21529.

Yu, G. *et al.* (2012) ‘ClusterProfiler: An R package for comparing biological themes among gene clusters’, *OMICS A Journal of Integrative Biology*, 16(5), pp. 284–287. doi: 10.1089/omi.2011.0118.

Yu, Y. *et al.* (2014) ‘Inhibition of KIF22 suppresses cancer cell proliferation by delaying mitotic exit through upregulating CDC25C expression’, *Carcinogenesis*, 35(6), pp. 1416–1425. doi: 10.1093/carcin/bgu065.

Zerbi, V. *et al.* (2018) ‘Dysfunctional Autism Risk Genes Cause Circuit-Specific Connectivity Deficits With Distinct Developmental Trajectories’, *Cerebral Cortex*, 28, pp. 2495–2506. doi: 10.1093/cercor/bhy046.

Zhong, S. *et al.* (2018) ‘A single-cell RNA-seq survey of the developmental landscape of the human prefrontal cortex’, *Nature*. Nature Publishing Group, 555(7697), pp. 524–528. doi: 10.1038/nature25980.

Ziats, M. N., Edmonson, C. and Rennert, O. M. (2015) ‘The autistic brain in the context of normal neurodevelopment’, *Frontiers in Neuroanatomy*. Frontiers Research Foundation, 9(AUGUST). doi: 10.3389/fnana.2015.00115.